Research Summary

Giles Hooker

My research interests lie principally in the fields of machine learning, statistical inference for dynamic systems models, and functional data analysis. I have also contributed to the literature on item response theory and continue to work in disparity-based inference. Although I have completed and current projects specific to each of these fields, many of my interests lie in the intersection of them. A detailed account of my achievements and future interests is given in the following pages, but I present a summary below.

Machine Learning methods (support vector machines, regression trees and others) are typically aimed at providing accurate predictions of some quantity. However, these prediction functions are usually difficult to interpret and are rarely associated with a specific probability model. I am interested in the interface between these type of methods and more classical statistical inference. This encompasses both statistical inference about structure that a prediction function tries to mimic, uncertainty in the explanations given about predictions resulting from machine learning, and the use of machine learning methods in statistical inference, incorporating additive models and latent variables.

Dynamical Systems: I am interested in the interface between statistical methods and nonlinear dynamics in the form of differential equations and stochastic models. These are motivated by applications in ecology, epidemiology and immunology. I have developed tools to estimate and provide inference about parameters in ODE models that are robust to stochastic disturbances. I have also produced methods for model criticism for dynamic systems models and the use of control-theoretic methods in designing model inputs so that experiments provide maximal information about parameters of interest. Longer-term goals include a study of robust methods of inference in these systems and experimental design aimed at model improvement.

Functional Data Analysis: I have a number of projects motivated by applications in ecology, vehicular emissions, remote sensing and fMRI data. My recent projects involve quantifying curvature in single index models. This stems from collaborations with ecologists who wish to understand the effect of environmental variability (quantified as an unknown combination of historical weather data) on biodiversity, where Jensen's inequality (applied to species' response to environment) plays an important role in allowing competing species to co-exist. I also have current work on applications in human biomechanics.

I have also worked on methods and models for latent functional random variables; on model selection in functional regression: if a functional covariate is used, what parts of the function are relevant? and on functional convolution models. These all have poorly characterized identifiability conditions; a similar current project examines selecting which derivative of a functional covariate should be used.

Research Statement

Giles Hooker

Detailed Description of Research Directions and Achievements

Throughout the following "under review" refers to manuscripts under review or currently in revision, "current projects" indicates substantial methodological development and experimentation has already been undertaken; where manuscripts are near submission, this has been indicated.

Highlight Achievements and Contributions

- Read paper in JRSS(B), publications in AISTATS, American Naturalist, AOAS, Bernoulli, Biometrics, Ecology Letters Entropy, FODS, KDD, JASA, JCGS, JMLR, J. Royal Society Interface, JRSSB, Nature Methods, Physical Letters Review, Psychometrika.

- PI or co-I on 14 successful grants, (total funding: $6,153,594, lead or co-lead on 5 with funding $1,825,657); including NSF CAREER award for "Diagnostics and Experimental Design in Nonlinear Dynamics."

- Published 2 books: "Functional Data Analysis in R and Matlab" and "Dynamic Data Analysis: Modeling Data with Differential Equations" with J. O. Ramsay.

- Organized 6 workshops on Statistical Methods for Nonlinear Dynamics, and Inference in Machine Learning; Associate Editor, Electronic Journal of Statistics, Journal of the American Statistical Association, Journal of the Royal Statistical Society; Program committee International Workshop on Statistical Modeling; Invited participant in 2 SAMSI programs; Short courses on Functional Data Analysis for IWSM 2010 and JSM 2015; Introductory Overview Lecture, JSM 2020; Board member of STAT.org.

- Developer on R and Matlab packages for functional data analysis, inference in dynamic systems models and inference with Random Forests.

Machine Learning

My interest in machine learning dates to my PhD dissertation with Jerry Friedman. I have since developed these ideas in collaboration with the Cornell Laboratory of Ornithology and more recently at the Cornell Weil College of Medicine and with the Department of Actuarial Studies at the University of New South Wales. While my early faculty career focussed on the dynamical systems models described in the next section, over the last few years I have returned my focus here as the machine learning community has become more interested in diagnostic tools and stemming from some of my own theoretical development.

Much of machine learning is devoted to the problem of prediction from high-dimensional covariates without postulating parametric forms for the relationship between the predicted outcome and the predictions. Many successful methods have been developed for this purpose, however they result in algebraically complex models that share properties of sacrificing interpretability in favor of predictive accuracy and having little by the way of formal uncertainty quantification. My interests have been in developing diagnostics to "understand" the resulting models, both from a global point of view and more recently in providing explanations for specific predictions, and in incorporating this very powerful class of tools as components in commonly used statistical models and into statistical inference.

My PhD dissertation in 2004 focussed on providing a global understanding of black box functions. This included trying to find answers to the following questions:

- Which covariates are important?
- What is the structure of an additive model that best represents the learned prediction function?
- How can we quantify the statistical stability of conclusions about these structures?

An important focus of my research was to avoid basing answers to these questions on the behavior of the learned function in regions of covariate space with little or no data. To that end I have developed extensions of the functional ANOVA to be based on non-independent distributions of the covariates (Hooker 2004b, 2007, Hooker *et. al.* 2021) as well as tree-based methods for estimating distributions of data (Hooker 2004a) and methods to improve the extrapolation performance of machine learning methods (Hooker and Rossett, 2011).

At Cornell I have developed a collaboration with the Laboratory of Ornithology in particular in their citizen science projects where data sets of millions of records from amateur bird watchers along with thousands of covariates have been collected. This is the first time that data on continent-wide bird abundance as been available and motivates the problem of "hypothesis generation": variables affect bird abundance? What combinations of variables have important non-additive joint effects? This project has resulted in two publications in the ecological literature (Fink *et. al.* 2009, Kelling *et. al.* 2010).

My most recent projects involve the integration of machine learning methods – particularly those based on bagging or other averages of random trees – with more common statistical practice. To this end, a former PhD student, Lucas Mentch, and developed a central limit theorem for the predictions of subsampled trees (Mentch and Hooker, 2015). The predictions of random forests are normally distributed, with a variance that we can estimate at no additional cost! This, in turn, will allow us to construct formalized tests of variable importance, interaction between variables as well as putting confidence intervals around summaries of the prediction function such as partial dependence plots (Mentch and Hooker 2017). This represents the first time that classical predictive tools can be combined with formalized statistical inference and opens an entirely new field of inferential techniques.

My current projects follow on from these; my former student, Yichen Zhou focussed on developing a central limit theorem for a particular version of gradient boosting (Zhou and Hooker, 2022). While we do not have this result in the generality that we would like – we have very restrictive assumptions on how trees are built – there is good evidence that it applies more broadly. We are further able to extend this to incorporate parametric effects within boosting. That is, we can produce models in which some components can be modeling using traditional statistical tools while we can rely on tree-based methods to account for components where specific understanding is not important and where the would be wasted. A particular example from Cornell's Laboratory of Ornithology is in dividing bird detectability (modeled parametrically as a function of effort variables) and a non-parametric estimate of species distributions. Another student, Indrayudh Ghosal, has examined using random forests themselves within a boosting algorithm (Ghosal and Hooker, 2021). In a regression framework, a single boosting step almost universally improves test set performance, and we are able to extend the central limit theorem to cover any finite series of steps. Zhengze Zhou developed improved variance estimates and in so doing extended the range of models for which the central limit theorem above holds (Zhou *et. al.* 2021). More recently, Ghosal *et. al.*, 2022 develop the Infinitesimal Jacknife as a generic tool for variance estimation that can be used for uncertainty quantification when comparing or combining differing models.

My current work examines explanations from a statistical perspective: if the explanations are not stable, what use are they? In the context of *model distillation* – in which we train an interpretable model to mimic a black box – we show in Zhou *et. al.*, 2018 that using decision trees to generate explanations results in explanations that can be very different each time the tree is generated. Such explanations can be stabilized, but doing so requires generating orders of magnitude more data than is usually used. Zhou and Hooker, 2021a repeated this analysis for LIME reaching similar results, while current work with Berkeley students Yunzhe

Zhou and Jenny Xu extends this to a generic framework for structure discovery in distillation method. In ongoing projects, I am developing methods for *counterfactual explanations* – What would I have to change to get a different prediction? – that in particular seek to identify and incorporate how feasible suggested actions are, along with Berkeley student Alex Assemota, stabilizing SHAP values and extracting more relevant information about their importance with Jeremy Goldwasser, and investigating social disparities in which explanations are given, specifically focussing on insurance premiums with colleagues at UNSW.

I have also returned to global diagnostics as a means of assessing fairness in machine learning. This project has particularly been undertaken with my student Sarah Tan and with Rich Caruana where we examine several data sets in which we have both a prediction tool and separately collected data (eg the COMPAS) and can attempt to examine what patterns the original tool uses and whether these are also evident in data (Tan *et. al.* 2018). Our work is based around generalized additive models (eg Lou *et. al.* 2012)and includes inference based on our results above, and can also be used to understand the success (or failure) of transfer learning. (Lengerich *et. al.*, 2020) follows on from this in adding interaction effects. Sarah Tan also worked with me on prototype methods (Tan *et. al.*, 2020) as another means of making nearest-neighbours-like methods interpretable.

I think there is a huge range of opportunities (and equally as many pitfalls) in bringing the very powerful tools in machine learning into statistical inference. Future work includes extensions of uncertainty quantification results to deep learning, screening tools for variable importance and including latent variable models. I am also keenly interested in applications of these techniques and look forward to expanding my range of collaborations.

Grants

- NSF TRIPODS 1740882 2017: "Data Science for the Social Good", $1,497,238, co-PI with K. Weinberger, J. Kleinberg, S. Strogatz and D. Schmoys.
- NSF DMS-1712554, 2017: "Statistical Inference Using Random Forests and Related Methods", $335,078, PI with L. Mentch.
- NVIDIA Hardware Grant: "Linking Convolutional Neural Networks, Random Forests and Statistical Inference", PI.
- NSF DEB-1353039 2014: "Integral Projection Models for Populations in Varying Environments: Construction and Analysis", $652,847, co-PI with S. Ellner, P. Adler and R. Snyder.
- NIH R03DA036683, 2014: "Shortening the SOAPP-R: Computer-based opioid risk assessment" $160,579, co-PI with M. Finkelman.

- NSF CDI Type II, 2011: “Bircast: Novel Machine Learning Methods for Understanding Continent-Scale Migration”, $1,217,895. co-I with S. Kelling, and T. Diettrich.

Peer-Reviewed Publications

- Tan, Sarah, **Giles Hooker**, Paul Koch, Alberto Gordo and Rich Caruana, 2023, “Considerations When Learning Additive Explanations for Black-Box Models”, *Machine Learning*, in press.
- Zhou, Yichen, Zhengze Zhou and **Giles Hooker**, 2023, “Approximation Trees: Statistical Stability in Model Distillation”, *Data Mining and Knowledge Discovery*, in press.
- Zhou, Yichen and **Giles Hooker**, 2022, “Boulevard: Regularized Stochastic Gradient Boosted Trees and Their Limiting Distribution”, *Journal of Machine Learning Research*, 23(183):1-44.
- Zhou, Yichen and **Giles Hooker**, 2022, “Tree Boosted Varying Coefficient Models”, *Data Mining and Knowledge Discovery*, in press.
- Hooker, Cliff, Claire Hooker and **Giles Hooker**, 2022, “Expertise, a Framework for our Most Characteristic Asset and Most Basic Inequality”, *Spontaneous Generations*, 10(1) “The Revolt Against Expertise” pp 27-35.
- Zhou, Zhengze, Lucas Mentch and **Giles Hooker**, 2021, “V-Statistics and Variance Estimation” *Journal of Machine Learning Research*, 22(287):1-48.
- **Giles Hooker**, Lucas Mentch, and Siyu Zhou, 2021, “Unrestrected Permutation Forces Extrapolation: Variable Importance Requires at least One More Model, or There is No Free Variable Importance”, *Statistics and Computing*, 31(6):82.
- Ghosal, Indrayudh and **Giles Hooker**, 2021, “Boosting Random Forests to Reduce Bias; One-Step Boosted Forest and its Variance Estimate”, *Journal of Computational and Graphical Statistics*, 30(2):493-502.
- Zhou, Zhengze, and **Giles Hooker**, 2021, “Unbiased Measurement of Feature Importance in Tree-Based Methods”, *Transactions in Knowledge Discovery and Data Mining*, 15(2):1-21.
- Coleman, Tim, Lucas Mentch, Daniel Fink, Frank La Sort, **Giles Hooker**, Wesley Hochachka and David Winkler, 2020, “Statistical Inference on Tree Swallow Migrations”, *Journal of the Royal Statistical Society, Series C*, 69(4):973-989.

- J. Wen, P. Köhler, G. Duveiller, N.C. Parazoo, T.S. Magney, **G. Hooker**, L. Yu, C. Y. Chang, and Y. Sun, 2020 "Generating a Long-Term Record of High-Resolution Global Solar-Induced Chlorophyll Fluorescence (SIF) by Harmonizing Multiple Satellite Instruments: A Case Study for Fusing GOME-2 and SCIAMACHY", *Remote Sensing of Environment*, 239:111644.

- Lengerich, Ben, Sarah Tan, Chun-Hao Chang, **Giles Hooker** and Rich Caruana, 2020, "Purifying Interaction Effects with the Functional ANOVA:An Efficient Algorithm for Recovering Identifiable Additive Models", *AISTATS*.

- Tan, Sarah, Matvey Soloviev, **Giles Hooker**, and Martin T. Wells 2020, "Tree Space Prototypes: Another Look at Making Tree Ensembles Interpretable", *FODS*.

- **Hooker, Giles** and Lucas Mentch, 2018, "Bootstrap Bias Corrections for Ensemble Methods" *Statistics and Computing*, 28(1):77-86.

- **Hooker, Giles** and Cliff Hooker, 2018, "Machine Learning and the Future of Realism", *Spontaneous Generations: A Journal for the History and Philosophy of Science*, 9(1):174-182.

- Mentch, Lucas and **Giles Hooker**, 2017, "Formal Hypothesis Tests for Additive Structure in Random Forests", *Journal of Computational and Graphical Statistics*, 26(3):589-597.

- Tan, Sarah, Rich Caruana, Giles Hooker and Yin Lou, 2018, "Distill-and-Compare: Auditing Black-Box Models Using Transparent Model Distillation", AAAI/ACM Artificial Intelligence, Ethics, and Society 2018.

- Kang, Keegan and **Giles Hooker**, 2017, "Random Projections with Control Variates", *Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods.*

- Kang, Keegan and **Giles Hooker**, 2016, "Improving the Recovery Of Principal Components with Semi Deterministic Random Projections", *Proceedings of the 50th Annual Conference on Information Science and Systems.*

- Keegan and **Giles Hooker**, 2016, "Block Correlated Deterministic Random Projections", *Proceedings of the 6th Conference on Computational Mathematics, Computational Geometry and Statistics.*

- Mentch, Lucas and **Giles Hooker**, 2015, "Quantifying Uncertainty in Random Forests via Confidence Intervals and Hypothesis Tests", *Journal of Machine Learning Research*, 17(3):1-41

- Grinspan, Zachary, M., JS Shapiro, Erika L. Abramson, **Giles Hooker**, Rainu Kaushal and Lisa M. Kern, 2015, "Predicting Frequent ED Use By People with Epilepsy with Health Information Exchange Data", *Neurology,Neurology*, 85(12):1031-1038.

- **Giles Hooker**, 2013, "A review of Boosting: Foundations and Algorithms by Schapire and Freund", *Journal of the American Statistical Association*, 108(502):750-754..

- Lou, Yin, Rich Caruana, Johannes Gehrke and **Giles Hooker**, 2013, "Accurate Intelligible Models with Pairwise Interactions", *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

- Robert D. Gibbons, **Giles Hooker**, Matthew D. Finkelman, David J. Weiss, Paul A. Pilkonis, Ellen Frank, Tara Moore and David J. Kupfer, 2013, "Computerized Adaptive Diagnosis of Depression Using the CAD-MDD", *Journal of Clinical Psychiatry*, 74(7): 669-674.

- **Giles Hooker** and James Ramsay, 2012, "Learned-Loss Boosting", *Computational Statistics and Data Analysis*, 56:3935-3944.

- **Giles Hooker** and Saharon Rosset, 2012, "Prediction-Focussed Regularization Using Data-Augmented Regression", *Statistics and Computing*, 1:237-248.

- Daniel Fink, Wesley M. Hochachka, Benjamin Zuckerberg, David W. Winkler, Ben Shaby, M. Arthur Munson, **Giles Hooker**, Mirek Riedewald, Daniel Sheldon and Steve Kelling, 2010, "Spatiotemoral Exploratory Models for Broad-scale Survey Data", *Ecological Applications*, 20:2121-22147.

- Steve Kelling, Wesley M. Hochachka, Daniel Fink, Mirek Riedewald, Rich Caruana, Grant Ballard and **Giles Hooker**, 2009,"Data Intensive Science: A New Paradigm for Diversity Studies". *Biosciences*, 59:613-620.

- **Giles Hooker**, 2007. "Generalized Functional ANOVA Diagnostics for High Dimensional Functions of Dependent Variables". *Journal of Computational and Graphical Statistics*. 16:709-732.

- **Giles Hooker**, 2004a. "Diagnosing Extrapolation: Tree-Based Density Estimation". *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

- **Giles Hooker**, 2004b. "ANOVA Diagnostics for Black Box Functions". *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

- **Giles Hooker** and Matthew Finkelman, 2004. "Sequential Analysis for Learning Modes of Browsing". *WEBKDD 2004: Proceedings of the Sixth International Workshop on Knowledge Discovery from the Web.*

Commentary, Abstracts and Software

- **Hooker, Giles** and Lucas Mentch, 2019 "Please Stop Permuting Features: Explanations and Alternatives"

- Seto, Skyler, Sarah Tan, **Giles Hooker**, Martin T. Wells and John S. Johnson, 2017, "A Double Parametric Bootstrap Test for Topic Models", *NIPS 2017 Symposium on Interpretable Machine Learning.*

- Tan, Sarah, Rich Caruana, **Giles Hooker** and Yin Lou, 2017, "Detecting Bias in Black-Box Models Using Transparent Model Distillation", *NIPS 2017 Symposium on Interpretable Machine Learning.*

- **Hooker, Giles** and Lucas Mentch, 2016, "Comments On: A Random Forest Guided Tour", *TEST* 25(2):254-260.

- Tan, Hui Fen, **Giles Hooker** and Martin T. Wells, 2016, "Tree Space Prototypes: Another Look at Making Tree Ensembles Interpretable", *NIPS 2016 Workshop on Interpretable Machine Learning in Complex Systems.*

- Tan, Hui Fen, **Giles Hooker** and Martin T. Wells, 2016, " Probabilistic Matching: Incorporating Uncertainty to Correct for Selection Bias", *NIPS 2016 Causal Inference Workshop.*

- "SuRFIn: Subsampled Random Forest Inference", `R` package in preparation.

- "Adaptive-Loss Boosting", Matlab routines.

Under Review

- Zhou, Yunzhe, Peiru Xu and **Giles Hooker**, "Stability and Generic Model Distillation"

- Ghosal, Indrayudh, Yunzhe Zhou and **Giles Hooker**, "The Infinitesimal Jacknife and Combinations of Models".
- Ghosal, Indrayudh and **Giles Hooker**, "Generalized Boosted Forests".
- "Price Dynamics on Amazon Marketplace: A Machine Learning Approach", with Sharmistha Sikdar
- Sarah Tah, Rich Caruana, **Giles Hooker** and Yin Lou, "Auditing Black-Box Models Using Transparent Model Distillation With Side Information"

Current Projects

- "Obtaining the Feasibility of Counterfactual Explanations from Longitudinal Data", with Alex Assemota
- "Getting more from Shapley Values" with Jeremy Goldwasser
- "Density Estimation with BART", with Ed George and Rob McCullogh
- "Targeted Machine Learning for Integral Projection Models" with Yunzhe Zhou
- "Theory of Random Forests: A Review" with Erwan Scornet for *Annual Reviews*

Statistics for Dynamic Systems

Much of my research focus has been directed towards developing inferential methods using nonlinear dynamical models. This work attempts to bridge the fields of applied mathematics and statistics. While statistical models have traditionally been phenomenological in nature - based on linear or nearly-linear processes - applied mathematics has tended to use nonlinear models, particularly systems of ordinary differential equations, to model and explain physical systems. The use of dynamical systems has a number of appealing facets: they are typically mechanistic; every component having a clear causal interpretation that can often be derived from physical first principles, their associated bifurcation theory also provides a powerful, causal explanation for their qualitative behavior. However, these models have gained little statistical attention, and few methods exist to compare them to data. This is for two main reasons: (i) the numerical difficulties involved in working with these models, and (ii) their dependence on simplified, deterministic dynamics results in poor quantitative agreement with observed data from systems that cannot be as well controlled as the models suggest.

My research in this area can be classed into three problems

1. Estimation and inference for parameters in ODEs based on smoothing methods to provide robustness against model miss-specification and stochastic disturbances. These methods were developed in conjunction with Jim Ramsay and his students. These were published as a read paper in JRSS(B) (Ramsay *et. al.* 2007, below). I have written accompanying Matlab software and have since developed an R package: `CollocInfer` as part of a program at the National Center for Ecological Analysis and Synthesis. I have further refined these methods and provided improved confidence intervals and smoothing parameter selection methods as part of a study on the dynamics of measles epidemics in Ontario (Hooker *et. al.* 2011).

   A future project focusses on connections between these methods and inference in models given explicitly as stochastic disturbances to explicitly describe the robustness properties of profiling, and to examine the effect of stochasticity on the precision of parameter estimates.

2. Diagnostics for model lack-of-fit. My focus here has been on methods for model improvement in which the model defines rates of change which cannot be directly compared to observations. A first paper in this project was published in Biometrics (Hooker, 2009) which provided lack-of-fit tests for ODEs along with graphical diagnostics for model improvement. A more recent paper in AOAS (Hooker and Ellner, 2015) focussed on

(a) Distinguishing lack of fit due to purely stochastic sources, independent of the observed system from lack of fit in the specification of the systems dynamics.

(b) Distinguishing lack of fit due to miss-specification of functional forms in and ODE from lack of fit due to missing components of the state vector.

This creates a hierarchy of sources of lack of fit: stochastic disturbance, function miss-specification, system miss-specification that can be investigated.

These are motivated by collaborations with mathematical and experimental ecologists where observed dynamics in a tightly controlled two-species system do not correspond to theoretical predictions. It is proposed that real-time evolution is responsible for this lack of fit which can be modeled as an additional species (ie state variable) in the system.

The proposed diagnostic tests are based on tools developed in the literature on attractor reconstruction and make use of machine learning methods to assess the predictive accuracy of including lagged quantities in forecasting future dynamical behavior.

3. Experimental Design. Within experimental ecology, as well as in many other application areas, there are frequently inputs into dynamic systems of interest that can be controlled by the experimenter. Within the context of experimental ecology, algae and rotifers are kept in a contained environment to which nutrients are continuously added at a controllable rate.

   An important question is how these inputs should best be designed so as to provide most information about parameters of interest. An investigation into this was begun as part of the SAMSI program for Stochastic Dynamics where methods from control theory have been employed to provide optimal adaptive experimental design in the context of stochastic dynamic systems. Systems which have noisy and partial observations pose considerable numerical and theoretical challenges. This has resulted in two papers, Hooker, Lin and Rogers, 2015, demonstrated that within the context of exactly observed diffusion processes the problem could be approached within the framework of control theory. A paper with a former student (Thorbergsson and Hooker, 2018), employs these ideas within the framework of Partially Observed Markov Decision Processes.

   More recent work with Aurya Javeed (Javeed and Hooker, 2020) extends these ideas to the timing of observations when – as in the caste of counting numbers of rotifers in a sample – individual observations are expensive. More commercial application can apply to blood tests designed to examine absorbtion rates in pharmaco-kinetics.

4. Nonparametric models for differential equations. The use of machine learning tools to estimate differential equations has become increasingly popular, with a number of proposals to estimate the right hand side functions of differential equations without resorting to parametric effects. A current project with Berkeley students William Torous and Abhineed Agarwal looks at the way that the behaviors of system affect our ability to conduct nonparametric estimation for it. As a case in point, a system that exhibits cyclic behavior only generates data on the path of its limit cycle, and non-parametric estimates of its dynamics do not produce systems that also exhibit this qualitative behavior. We aim to provide tools that make observed paths at least locally stable. Further work will focus on inference about the stability properties of systems from nonlinear estimates of them.

5. Models for high-dimensional systems. Systems biology examines the interactions of hundreds to thousands of components of cellular functioning and high-throughput experiments are beginning produce time series of the resolution to allow dynamic models to be fit. A first task for this is to discover network structure from these models.

   Some work has been undertaken along these lines assuming linear dynamics in which standard sparsity penalties can be applied within a two-stage method. With a former undergraduate student, Yujia Zhang, I have begun to extend these to the saturation effects and interactions that are typically found in biological models making this approach far more realistic.

Beyond methodological research, I have been active in developing a community of researchers in a still very young field of statistics. In particular, I have helped to organize workshops on statistical inference in nonlinear dynamics at the Centre de Reserches Mathematiques in 2007, the Pacific Institute of Mathematical Sciences in 2009 and the Mathematical Biology Institute in 2012 as well as leading a successful workshop at the Banff International Research Station in July 2014. I have also been an invited participant in two SAMSI programs mentioned including as a visitor and working group leader in the program on the Analysis of Object Data. I have written a book with Jim Ramsay detailing methods on fitting ordinary differential equations, and have substantially completed work on a much expanded second edition, now under contract with Springer, that will include stochastic models. This is also supported under the CAREER award listed below.

Grants

- NSF DEB-1353039, 2014: "Integral Projection Models for Populations in Varying Environments: Construction and Analysis", $652,847, co-PI with S. Ellner, P. Adler and R. Snyder.

- NSF DEB-125619, 2013: "Effects of Rapid Consumer Evolution on Community Dynamics: predictions and Tests in a (Nearly) Natural Food Web", $200,000, co-PI with S. Ellner and N. Hairstone.
- NSF DMS-1053252 CAREER, 2010: "Diagnostics and Experimental Design for Nonlinear Dynamics", $400,000. PI.
- NSF DEB-0813743, 2008: "Rapid Evolution and the Dynamics of Complex Ecological Communities", $539,957, co-PI with S. Ellner, L. Jones and N. Hairstone.
- Hatch NYC-150446, 2007: "Experimental Design for Nonlinear Processes in Agriculture", $30,000, PI.

Book

- Ramsay, James O. and **Giles Hooker**, 2017, "Dynamic Data Analysis: Modeling Data with Differential Equations", Springer.

Peer-Reviewed Publications

- Javeed, Aurya, and **Giles Hooker**, 2020, "Timing Observations of Diffusions", *Statistics and Computing*, 30:405-417.
- Thorbergsson, Leifur and **Giles Hooker**, 2018, "Experimental Design for Partially Observed Markov Decision Processes", *Journal of Uncertainty Quantification*, 6(2):549-567.
- **Hooker, Giles**, James O. Ramsay and Luo Xiao, 2016, "CollocInfer: Collocation Inference in Differential Equation Models", *Journal of Statistical Software*, 75(2)
- **Hooker, Giles** and Stephen P. Ellner, 2015, "Goodness of Fit Diagnostics in Nonlinear Dynamics: Mis-specified Rates or Mis-specified States?". *Annals of Applied Statistics*, 9(2):754-776.
- **Hooker, Giles**, Kevin K. Lin and Bruce Rogers, 2015, "Control Theory and Experimental Design in Diffusion Processes", *Journal of Uncertainty Quantification*, 3(1):234-264.
- Hiltunen, Teppo, Nelson G. Hairstone, **Giles Hooker**, Laura E. Jones and Stephen P. Ellner, 2014, "A newly discovered role of evolution in previously published consumer-resource dynamics", *Ecology Letters*, 17(8):915-923.

- Hiltunen, Teppo, Stephen P. Ellner, **Giles Hooker**, Laura E. Jones, Nelson G. Hairston, 2014, "Eco-evolutionary Dynamics in a Three-Species Food Web with Intraguild Predation: Intriguingly Complex" in *Advances in Ecological Research, Vol. 50 – Eco-Evolutionary Dynamics*, Jordi Moya-Laraño, Jennifer Rowntree and Guy Woodard, Editors.

- David Campbell, **Giles Hooker** and Kim McAuley, 2012, "Parameter Estimation in Differential Equation Models with Constrained States", *Journal of Chemometrics*, 56:322-332.

- **Giles Hooker**, Stephen. P. Ellner, Laura de Vargas Roditi and David J. D. Earn, 2011, "Parameterizing State-space Models for Infectious Disease Dynamics by Generalized Profiling: Measles in Ontario", *J. Royal Society Interface*, 8:961-975.

- **Giles Hooker**, 2010, "Comments on: Dynamic Relations for Sparsely Sampled Gaussian Processes", *TEST*, 19, 50-53.

- Ercan Atam and **Giles Hooker**, 2010, "An Identification-based State Estimation Method for a Class of Nonlinear Systems". *J. Systems and Control Engineering*, 224:349-359.

- **Giles Hooker**, 2009, "Forcing Function Diagnostics for Nonlinear Dynamics". *Biometrics*, 65:928-936.

- James O. Ramsay, **Giles Hooker**, David Campbell and Jiguo Cao, 2007. "Parameter Estimation for Nonlinear Differential Equations: A Smoothing-Spline Approach". *Journal of the Royal Statistical Society, Series B*, 69:741-796.

Software

- "Collocation Inference in Nonlinear Stochastic Dynamics", `R` Package and Manual.

- "Smoothing Methods for Nonlinear Dynamics". Manual and MATLAB Software.

Technical Reports

- **Giles Hooker** and Stephen P. Ellner, 2010, "On Forwards Prediction Error", Technical Report BU-1679-M, Department of Biological Statistics and Computational Biology, Cornell University.

- **Giles Hooker** and Lorenz T. Biegler, 2007. "IPOPT and Neural Dynamics: Tips, Tricks and Diagnostics", Technical Report BU-1676-M, Department of Biological Statistics and Computational Biology, Cornell University.

- **Giles Hooker**, 2007. "Theorems and Calculations for Smoothing-Based Profiled Estimation of Differential Equations". Technical Report BU-1671-M, Department of Biological Statistics and Computational Biology, Cornell University.

Under Review

- Sharmistha Sikdar and **Giles Hooker**, "A Hidden Semi-Markov Model of Customers" Multi-channel Engagement Dynamics"

Current Projects

- "Variable Selection and Parameter Estimation of High-dimensional Michaelis-Menten Models" with Yujia Zhang

- "Gradient Matching, Robustness, and the Linear Noise Approximation"

- "A New Understanding of Principal Differential Analyais" with Edward Gunning

- "Nonparametric ODEs and Qualitative Behavior" with Abhineed Agarwal and William Torous

Functional Data Analysis

My interests in functional data analysis stem from my post-doctoral fellowship undertaken under the direction of Jim Ramsay where we employed FDA methods for inference in dynamic systems. I strongly believe in providing usable software to accompany methodological development. To that end, I have worked closely with Jim Ramsay and Spencer Graves in the development of the `fda` library in `R` where I attempt to incorporate new methodology as it is developed. This has also resulted in a book on the practical implementation of functional data analysis which was one of Springer's best sellers at the JSM, 2010. I have also given a short course to the International Workshop on Statistical Modeling in June 2010 and the Joint Statistical Meetings, 2015.

Active collaborations have also motivated the development of new methodology in functional data analysis itself. Specifically:

1. Convolution models represent a continuous-time generalization of distributed lag models in time series. These models are motivated by applications in modeling vehicle exhaust emissions and represent the exhaust as a functional response that depends on an integral over the short-term past of covariates. Asencio *et. al.* 2014) presents an initial paper presenting bootstrap methods and examining identifiability.

   As an offshoot from this model the length of the convolution – how far into the past do the covariates have an effect – represents an important inferential question here as well as in many other applications of FDA. As part of the SAMSI program on the Analysis of Object Data, I have an active project in "Domain Selection" – providing inference about what parts of a functional covariate influence the response of interest. An initial paper with Peter Hall, (Hall and Hooker, 2016) was published in JRSS(B).

2. Modeling functional covariance. This project was motivated by satellite image data that provides measures of "greenness" in regions of the globe and stems from my grant with Mark Friedl and Surajit Ray at Boston University. Spatial and temporal variation is important for understanding ecosystem responses to climate change and for "gap-filling" missing data. A new factor rotation to improve the interpretation of functional principal components analysis was published in AOAS (Liu *et. al.*. 2012); connections to Maximal Autcorrelation Functions were studied in Hooker and Roberts, 2016; tests for isotropy are given in Liu *et. al.*. 2017.

   A further interest from this project has been on Bayesian estimation of covariance surfaces, which has became the topic of my former student Cecilia Earls. While there has been some interest in estimating covariance surfaces for functional data, there has been little Bayesian attention given to the problem. A starting point here is to extend classical multivariate methods to Gaussian processes. The Bayesian framework is attractive here because it allows the study of latent functional processes that occur in registration, with sparsely observed covariates and in structural equation models. Three papers (Earls and Hooker, 2014, 2017a,b) came out of this project, developing progressively more complex models to incorporate variation.

3. Functional Generalized Additive Models (FGAMs). These are extensions of generalized additive models to functional data was the main focus for Matthew McLean, a PhD student co-advised by myself and David Ruppert. There have been three parts to this work: the development of the model and smoothing-based estimation for it (McLean *et. al.* 2014a), a variational Bayes approach when covariate functions are sparsely observed, and a test of fit for the functional linear model using FGAMs as a more complex alternative (McLean *et. al.* 2014b).

4. Functional Single Index Models, and Ecological Inference. A long-standing problem in mathematical biology is the maintenance of biodiversity. How is it that species that use the same resources can coexist when standard evolutionary theory suggests that one should be out-competed? One mechanism that allows for coexistence is in differing responses to environmental variability: in particular if one species has an advantage in typical conditions but reacts very badly to poor conditions, a species that is less affected by these conditions, or takes better advantage of good times can still hold on.

   This is can be formalized in terms of the curvature of their *reaction norm*: their response to the environment. To assess this in real-world systems, we describe the environment via a functional linear model in which historical weather represents a functional covariate (see Teller *et. al.* 2016 for the first functional models employed in ecology). To assess curvature, my former student, Zi Ye, hs proposed estimating a nonparametric link function resulting in a functional single index model in which we have derived rates of convergence which are unsurprisingly slow (Ye and Hooker, 2020), but also direct estimation of the ecological effects of interest, which have substantially greater precision (Ye *et. al.* 2020) which we have now extended to generalized functional linear models (Ye *et. al.* 2021)

   This is part of a broader collaboration with both theoretical and mathematical ecologists on empirically quantifying the mechanisms maintaining biodiversity. This has included an analysis of specific effects (Adler *et. al.*, 2018, Tredennick *et. al.* 2018) and a general framework established in Ellner *et. al.*, 2018 (based partly on the functional ANOVA methods I developed in machine learning) that is rapidly becoming influential.

5. Selecting Derivatives in Functional Linear Models. One of features of functional data

that distinguishes it from multivariate data is the ability to take derivatives. This has been important in applications using spectral measurements of samples, for example. Nonetheless, formal tests for the appropriate derivative, and conditions where this can be identified, have not been developed. In a recent paper with Hanlin Shang, we develop these tests based on a simple use of integration by parts, but with somewhat subtle consequences (Hooker and Shang, 2022). Follow-up work will focus on more difficult problems of deciding with derivative of a function *response* to model; this raises identifiability issues, particularly in the context of the dynamic systems models discussed above.

6. Principal differential analysis and human biomechanics data. An aspect of functional data analysis that has received relatively little attention is relating derivatives to each other. Although this idea was first suggested by Jim Ramsay in 1996, the literature devoted to either studying or using these ideas is small and sparse. Together with Limerick University student Ed Gunning, I am revisiting these techniques when interpreted as providing estimates from an underlying nonlinear dynamical systems. Here we show that classical PDA estimates introduce bias, which we can correct, and that this perspective dictates certain properties of how you should center and register the data. These extensions are motivated by current work with the Australian Institute of Sport on human biomechanics in which we seek to reverse engineer control processes involved in running with the aim of improving recovery times from injury.

Grants

- NSF DEB-1933497, 2020: "Collaborative Research: A general approach to partitioning contributions from multiple drivers affecting individuals, populations, and communities", $750,000, co-PI with S. Ellner, P. Adler and R. Snyder.

- NSF DEB-1353039, 2014: "Integral Projection Models for Populations in Varying Environments: Construction and Analysis", $652,847, co-PI with S. Ellner, P. Adler and R. Snyder.

- NSF CMG-0934735, 2009: "Functional modeling of climate-ecosystem dynamics", $350,000, co-PI with S. Ray and M. Friedl.

Book and Short Course

- James O. Ramsay, **Giles Hooker** and Spencer Graves, 2009, "Functional Data Analysis in R and Matlab", Springer.

- Functional Data Analysis: Methods and Computing, *Joint Statistical Meetings*, Seattle, August 8, 2015.
- Functional Data Analysis, *International Workshop on Statistical Modeling*, Glasgow, July 3, 2010.

Peer-Reviewed Publications

- **Hooker, Giles** and Hanlin Shang, 2022, "Selecting the Derivative of a Functional Covariate in Scalar-on-Function Regression", *Statistics and Computing*, in press.
- Ye, Zi, **Giles Hooker** and Stephen P. Ellner, 2021, "Generalized Single Index Models and Jensen Effects on Reproduction and Survival", *Journal of Agricultural, Biological, and Environmental Statistics*, 26:492-512.
- Ye, Zi, **Giles Hooker** and Stephen P. Ellner, 2021, "Generalized Single Index Models and Jensen Effects on Reproduction and Survival", *Journal of Agricultural, Biological, and Environmental Statistics*, 26:492-512.
- Ye, Zi, **Giles Hooker** and Stephen P. Ellner, 2020, "The Jensen Effect and Functional Single Index Models: Estimating the Ecological Implications of Nonlinear Reaction Norms", *Annals of Applied Statistics*, 14(3):1326-12341.
- Warmenhoven, John, Norma Bargary, Dominik Liebl, Andrew Harrison, Mark Robinson, Edward Gunning and **Giles Hooker**, 2021, "PCA of Waveforms and Functional PCA: A Primer for Biomechanics", *Journal of Biomechanics*, 116:110106.
- Liu, Chong, Surajit Ray and **Giles Hooker**, 2017, "Functional Principal Components Analysis of Spatially Correlated Data", *Journal of Computational and Graphical Statistics*, 27(6):1639-1654.
- Earls, Cecilia and **Giles Hooker**, 2017b, "Combining Functional Data Registration and Factor Analysis", *Journal of Computational and Graphical Statistics*, 26(2):296-305.
- Earls, Cecilia and **Giles Hooker**, 2017a, "Adapted Variational Bayes for Functional Data Registration, Smoothing, and Prediction", *Bayesian Analysis*, 12(2):557-582.
- **Hooker, Giles** and Steven Roberts, 2016, "Maximal Autocorrelation Functions in Functional Data Analysis", *Statistics and Computing*, 26(5):945-950.

- Hall, Peter and **Giles Hooker**, 2016, "Truncated Linear Models for Functional Data". *Journal of the Royal Statistical Society, Series B*, 78(3):637-653
- McLean, Matthew W., **Giles Hooker** and David Ruppert , 2014b, "Restricted Likelihood Ratio Tests for Linearity in Scalar-on-Function Regression", *Statistics and Computing*, 25(5):997-1008.
- Matthew W. McLean, **Giles Hooker**, Ana-Maria Staicu, Fabian Schiepl and David Ruppert, 2014a, "Functional Generalized Additive Models", *Journal of Computational and Graphical Statistics*, 23(1):249-269.
- Earls, Cecilia, and **Giles Hooker**, 2014,"Bayesian Covariance Estimation and Inference in latent Gaussian Process Models", *Statistical Methodology*, 18:79-100
- Maria Asencio, **Giles Hooker** and H. Oliver Gao, 2014, "Functional Convolution Models", *Statistical Modeling*, 14(4):1-21.
- Chong Liu, Surajit Ray, **Giles Hooker** and Mark Friedl, 2012, "Functional Factor Analysis for Periodic Remote Sensing Data", *Annals of Applied Statistics*, 6:601-624.
- Marija Zeremski, **Giles Hooker**, Marla A. Shu, Emily Winkelstein, Queenie Brown, Don C. Des Jarlais, Leslie H. Tobler, Barbara Rehermann, Michael P. Busch, Brian R. Edlin, and Andrew H. Talal, 2011, "Induction of CXCR3- and CCR5-associated Chemokines during Acute Hepatitis C Virus Infection.", *Journal of Hepatology*, 55:545-553.

Software

- `fda` library in `R` and MATLAB for Functional Data Analysis.

Technical Reports

- **Giles Hooker**, 2013, "On the Identifiability of the Functional Convolution Model", Technical Report BU-1681-M, Department of Biological Statistics and Computational Biology, Cornell University.

Current Projects

- "Selecting the Derivative in Principal Differential Analysis",
- "A New Understanding of Principal Differential Analyais", with Edward Gunning

Disparity Methods

Disparity-based methods are a class of techniques that can broadly be described as providing statistical inference based on a comparison of a parametric probabilistic model to a non-parametric estimate of the distribution of observed data. They are remarkable in containing methods that achieve both robustness to outlying data as well as statistical efficiency when the parametric model is correct. These results have been known since the late 1970's, however much of their development has been confined to i.i.d. data with a few papers on linear regression.

My research program in this area has been undertaken in collaboration with Anand Vidyashankar and aims at developing disparity methods that extend to modern statistical models and methods for non-i.i.d. data. A first paper, on the use of disparity-based methods in Bayesian inference, has been published in *TEST*. A follow-up on the use of conditional density estimation in disparity methods appeared in *Bernoulli*. Recently, I continued this work with a paper with Yuefeng Wu in *Entropy* on the use of nonparametric Bayesian density estimation methods within disparity methods. Beyond these, we plan to investigate a variant on these methods in which we obtain a non-parametric estimate of the density of residuals (which therefore depends on parameters) and to use this within disparity methods. This has an important bias-reduction effect and the principal employed can be extended to a general class of multivariate models.

I have further plans to pursue these methods to model selection via sparsity penalties where we expect to be able to both retain oracle properties and demonstrate robustness. Extensions of disparity methods to machine learning have already been partially explored in the Learned-Loss papers listed under machine learning above. I am further interested in extending disparity methods to models with latent variables and in particular to providing robust inference with respect to the innovation distribution in time-series methods and stochastic differential equations.

Peer-Reviewed Publications

- Wu, Yuefeng and **Giles Hooker**, 2018, "Asymptotic Properties for Methods Combining Minimum Hellinger Distance Estimates and Bayesian Nonparametric Density Estimates", *Entropy*, 20(12):955.

- **Hooker, Giles**, 2016, "Consistency, Efficiency and Robustness of Conditional Disparity Methods", *Bernoulli*, 22(2):857-900

- **Hooker, Giles** and Anand N. Vidyashankar, 2014, "Bayesian Model Robustness via Disparities", *TEST*, 23(3):556-584.

- Markus Hegland, **Giles Hooker** and Stephen Roberts. 1999. "Finite Element Thin Plate Splines in Density Estimation". In *Computational Techniques and Applications, Proceedings of the Ninth Biennial Conference: CTAC99.* Journal of the Australian Mathematics S

Technical Reports

- **Giles Hooker** and Anand Vidyashankar, 2011, "Consistency and Efficiency of Conditional Disparity Methods", Technical Report BU-1670-M, Department of Biological Statistics and Computational Biology, Cornell University.
- **Giles Hooker**, 1999. "Developing a Spline-Smoothed Density". Technical Report, Research School of Information Sciences and Engineering, Australian National University.

Under Review

- Yuefeng Wu and **Giles Hooker**, "Hellinger Distance and Bayesian Non-parametrics: Hierarchical Models for Robust and Efficient Bayesian Inference".

Current Projects

- "Rotational Disparities: Robustness that Avoids High Dimensional Density Estimation"
- "Model Selection and Sparsity via Disparity Methods", with Anand Vidyashankar.
- "Bootstrap Bias Corrections for Multivariate Disparity Methods", with Anand Vidyashankar.

Item Response Theory

A final source of research problems has been from field of item response theory and educational testing. This grew out of a collaboration aimed at modeling web browsing behavior. There has been considerable interest in the educational testing literature in modeling tests as measuring multiple dimensions of ability simultaneously: the probability of giving a correct response to each question is modeled as having a logistic-type relationship with abilities.

We observed that using these models can result in the situation that it is possible for the estimate of an examinee's ability in one dimension to increase when a correct response is changed to incorrect. This possibility raises questions about test fairness - it could be in the examinee's best interest to give deliberately incorrect answers - and we labeled the phenomenon a "paradoxical result". Since making this observation we have

- Examined the prevalence of paradoxical results in real-world data and demonstrated it to be concerning.
- Established conditions under which paradoxical results occur for all common estimates of ability; for maximum likelihood estimates we showed that paradoxical results are unavoidable.
- Developed computationally-feasible methods to test whether a proposed test could yield a paradoxical result when combined with Bayesian estimation methods.
- Developed test-building algorithms to produce tests that are guaranteed to avoid paradoxical results when combined with Bayesian estimates.

This program has resulted in a series of four papers. Three of these have been published in Psychometrika – the top quantitative methods journal in Psychology – with a further publication in the Journal of Educational and Behavioral Statistics. These papers have gained some notice in the field - I am already aware of three further papers in Psychometrika that discuss my results and which have in turn stimulated my own thinking on the subject. Future projects are planned to provide conditions for the existence of paradoxical results in estimates of linear combinations of abilities as well as to investigate the question of whether avoiding paradoxical results is incompatible with the consistency and efficiency of ability estimates.

Peer-Reviewed Publications

- Matthew Finkelman, **Giles Hooker** and Zhen Wang, 2010, "Prevalence and Magnitude of Paradoxical Results in Multidimensional Item Response Theory", *J. Educational and Behavioral Statistics*, 35:744-761.

- **Giles Hooker**, 2010, “On Separable Tests, Correlated Priors and Paradoxical Results in Multidimensional Item Response Theory”, *Psychometrika*, 75:694-707.

- **Giles Hooker** and Matthew Finkelman, 2010, “Paradoxical Results and Item Bundles”. *Psychometrika*, 75:249-271.

- **Giles Hooker**, Matthew Finkelman and Armin Schwartzman, 2009, “Paradoxical Results in Multidimensional Item Response Theory”. *Psychometrika*, 74:419-442.

Technical Reports

- Matthew Finkelman, **Giles Hooker** and Zhen Wang, 2009, “Unidentifiability and Lack of Monotonicity in the Multidimensional Three-Parameter Logistic Model”. Technical Report BU-1678-M, Department of Biological Statistics and Computational Biology, Cornell University.

Collaborations in Ecology

Much of my work has been inspired by collaborations with ecologists, both with the Cornell Laboratory of Ornithology and in the Department of Ecology and Evolutionary Biology. These collaborations motivated my methodological work on both the interface of machine learning methods and statistics, and in statistical methods for models involving nonlinear dynamics. More recently, I have been part of a collaboration has spanned two recent grants and seeks to interface field data with models and questions that derive out of theoretical ecology. In particular, these projects seek to use models of ecological dynamics to understand a wide range of questions. Two examples of is are

- At what point in an individual's life do chance events have large impacts on their reproductive success? The answers to this have real implications for optimal management strategies, particularly for species conservation.
- If a species is driven close to extinction, will it tend to "bounce back"? Which ecological mechanisms most contribute to this maintenance of biodiversity? These have substantial implications for forecasting and managing the effects of climate change.

The types of mechanisms contributing to both answers include differing responses to environmental variability among competing species, release of competitive pressure from members of one's own species when rare, and the ability to "store" environmental good fortune, for example in seed banks. These mechanisms can be encoded and investigated in mathematical models, but the timescales or numbers involved do not allow us to directly observe them. Rather they are complex consequences of models obtained from studies that examine individuals growth, survival and reproduction, usually obtained from short-term surveys. A key task is then to translate modeling, and its attendant uncertainty, for individuals into conclusions about ecological processes which have real-world consequences.

The outputs of these collaborations range from relatively applied modeling problems, meta-analyses over databases of collected models, commentary and position papers on good statistical practice in developing models, and new statistical methodology in support of these questions: work with Zi Ye in functional data analysis (above) is a good example of this. My current methodological questions center around model selection and causal inference when the large-scale consequences of models is quantity of interest. I am actively developing extensions of targeted Machine Learning to these problems, something that also interfaces with my work with the Cornell Lab of Ornithology and will also apply understanding social disparities that result from the application of Machine Learning.

Peer-Reviewed Publications

- Hernández, Christina, Stephen P. Ellner, Peter B. Adler, **Giles Hooker** and Robin Synder, 2023, "An exact version of Life Table Response Experiment analysis, and the R package exactLTRE", *Methods in Ecology and Evolution*, in press.

- Ellner, Stephen P., Robin E. Snyder, Peter B. Adler and **Giles Hooker**, 2022, "Toward a "Modern Coexistence Theory" for Discrete and Spatial", *Ecological Monographs*, in press.

- Ellner, Stephen P., Peter B. Adler, Dylan Z. Childs, **Giles Hooker**, Tom E.X. Miller and Mark Rees, 2021, "A critical comparison of integral projection and matrix projection models for demographic analysis: Comment", *Ecology*, in press.

- Snyder, Robin, Stephen P. Ellner and **Giles Hooker**, 2021, "Time and chance: using age partitioning to understand how luck drives variation in reproductive success", *The American Naturalist*, 197(4):110-128..

- Tredennick, Andrew T., **Giles Hooker**, Stephen P. Ellner and Peter B. Adler, 2021, "A practical guide to selecting models for exploration, inference, and prediction in ecology", *Ecology*, 102.6:e03336.

- Ellner, Stephen P., Snyder, Robin E., Adler, Peter B. and **Giles Hooker**, 2019, "An Expanded Modern Coexistence Theory for Empirical Applications", *Ecology Letters*, 22(1):3-18.

- Tredennick, Andrew T., Brittany J. Teller, Peter B. Adler, 2018, **Giles Hooker** and Stephen P. Ellner, "Size-by-environment interactions: a neglected dimension of species' responses to environmental variation", *Ecology Letters*, 21(12):1757-1770.

- Ellner, Stephen P., Snyder, Robin E., Adler, Peter B. and **Giles Hooker**, 2018, "An Expanded Modern Coexistence Theory for Empirical Applications", *Ecology Letters*, 21(12):1757-1770.

- Adler, Peter B., Andrew Kleinhesselink, **Giles Hooker**, Brittany Teller and Stephen P. Ellner, 2018, "Weak interspecific interactions in a sagebrush steppe: evidence from observations, models, and experiments", *Ecology*, 99(7):1621-1632.

- Teller, Brittany J., Peter B. Adler, Collin B. Edwards, **Giles Hooker**, Robin E. Snyder and Stephen P. Ellner, 2016, "Linking demography with drivers: climate and competition", *Methods in Ecology and Evolution*, 7(2):171-183.

Consulting

In addition to theoretical research, I have occasionally been involved as a consulting statistician in miscellaneous research projects. This has resulted in a series of publications and grants listed below.

Grants

- CTSC Community Engagement Award, $20,000, "HCV Testing in NYC Commercial Sex Venues" co-I with K. Marks and D. Daskalakis.

Papers and Abstracts

- Warmenhoven, John, Andrew Harrison, Daniel Quintana, **Giles Hooker**, Edward Gunning and Norma Bargary, 2020, "Unlocking Sports Medicine Research Data while Maintaining Participant Privacy via Synthetic Datasets", *SportRxiv Preprints*
- **Giles Hooker**, Sophia Brumer, Chrisopher Zappa and Edward Monahan 2021, "Inferences to be Drawn from a Consideration of Power-Law Descriptions of Multiple Data Sets Each Comprised of Whitecap Coverage, WB, and 10-m Elevation Wind Speed Measurements", in *Recent Advances in the Study of Oceanic Whitecaps*, P. Vlahos and E. C. Monahan (Eds).
- Kilian, Nicole, Yongden Zhang, Lauren LoMonica, **Giles Hooker**, Derek Toomre, Choukri Ben Mamoun and Andreas. M. Ernst, 2020, "Trafficking and Localization of S-Palmitoylated Proteins in Plasmodium falciparum-Infected Erythrocytes", *BioEssays*, 1900145.
- Kilian, Nicole, Yongden Zhang, Lauren LoMonica, **Giles Hooker**, Derek Toomre, Choukri Ben Mamoun and Andreas. M. Ernst, 2019, "Trafficking and Localization of S-Palmitoylated Proteins in Plasmodium falciparum-Infected Erythrocytes", *under review.*
- Sinclair, David G., and **Giles Hooker**, 2019, "Sparse Inverse Covariance Estimation for High-throughput microRNA Sequencing Data in the Poisson Log-Normal Graphical Model". *Journal of Statistical Computation and Simulation*, 89(16)3105-3117.
- Sinclair, David G. and **Giles Hooker**, 2021, "An Expectation Maximization Algorithm for High-Dimensional Model Selection for the Ising Model with Misclassified States", *Journal of Applied Statistics*, in press.

- Goryaynov, Alexander, Nicole Kilian, Mark Lessard, Derek Doomre, James Rothman, **Giles Hooker** and Jörg Bewersdorf, 2018 "Assessing photodamage in live-cell STED microscopy", *Nature Methods*, 15:755-756.

- Monahan, Edward C, **Giles Hooker** and Christopher J. Zappa, 2015, "The Latitudinal Variation in the Wind-Speed Parameterization of Oceanic Whitecap Coverage: Implications for Global Modelling of Air-Sea Gas Flux and Sea Surface Aerosol Generation", *American Meteorological Society.*

- S.A. Jesty, S.W. Jung, J.M. Cordeiro, T.M. Gunn, J.M. Di Diego, S. Hemsley, B.G. Kornreich, **G. Hooker**, C. Antzelevitch, N.S. Moise, 2013, "Cardiomyocyte calcium cycling in a naturally occurring German shepherd dog model of inherited ventricular arrhythmia and sudden cardiac death", *Journal of Vetinary Cardiology* 15(1): 5-14.

- Anna Gelzer, Marcus L. Koller, Niels F. Otani, Jeffrey J. Fox, M. W. Enyeart, **Giles Hooker**, Mark L. Riccio, Carlo R. Bartoli and Robert F. Gilmour, 2008, "Dynamic Mechanisms for Initiation of Ventricular Fibrillation in vivo", *Circulation*, 118:1123-1129.

- Robert Norris, Jessica Ngo, Karen Nolan and **Giles Hooker**, 2005. "Volunteers are Unable to Properly Apply Pressure Immobilization in a Simulated Snakebite Scenario". *Journal of Wilderness and Environmental Medicine*, 16:16-21.

- Michael Shirts, Eric Bair, **Giles Hooker** and Vijay Pande, 2003. "Equilibrium Free Energies from Non-equilibrium Estimates Using Maximum Likelihood Methods". *Physical Letters Review*. 91(14):140601.

- **Giles Hooker** and Fuliang Weng, 2004. "Subset Selection in Large, Sparse Systems: An application of the Forward Stagewise approach to Natural Language Processing". Technical Report, Robert Bosch Corporation.