Precisely Wrong

How Pseudo-Absence Inflation Biases Parameter Estimates and Degrades Credible Interval Coverage in Bayesian Logistic Regression

Jeanne McClure, PhD

Masters of Data Science, Analytics and Modeling | University of West Florida

Advisors: Dr. Achraf Cohen & Dr. Samantha Seals

2026-01-01

The Methodological Choice Nobody Questions

When confirmed absences are unavailable, researchers generate **pseudo-absences** and fit logistic regression to a binary outcome (Guisan and Zimmermann 2000; Elith and Leathwick 2009).

The ratio of pseudo-absences to presences (PA:P) is an **analyst-controlled parameter** that varies widely across the published SDM literature — yet it is typically set by convention, not inferential justification.

What we know:

- Barbet-Massin et al. (2012) recommended large PA counts for regression-based SDMs based on **predictive accuracy** (True Skill Statistic)
- Northrup et al. (2022) warned that the available distribution likely contains used habitat **misclassified as absences** — a contamination problem that worsens as PA count grows

What we don't know:

- How does the PA:P ratio affect **parameter bias**?
- How does it affect **credible interval calibration**?
- At what point do your intervals stop telling the truth?

Why This Matters Now

Cox, Schmutz, and Seals (2025) applied presence-pseudo-absence logistic regression to loggerhead nesting data in the Florida Panhandle and cited Barbet-Massin et al. (2012) to justify a 1:1 PA:P ratio.

But that paper's recommendation for regression techniques is specifically a large-PA design — not a 1:1 balance.

The inferential consequences of this choice are well-grounded in theory:

- Firth (1993) showed MLE for logistic regression has $O(n^{-1})$ bias
- King and Zeng (2001) demonstrated systematic attenuation toward zero under low event prevalence
- Puhr et al. (2017) and Faghih et al. (2020) confirmed that 95% confidence intervals fail to achieve nominal coverage under severe imbalance
- In Bayesian settings, Gelman et al. (2004) and Northrup and Gerber (2018) showed credible intervals require the likelihood to dominate the prior — a condition that fails with small or rare-event samples

No prior study has quantified these effects across PA:P ratios in a Bayesian SDM.

Research Questions

RQ1: How does increasing the PA:P ratio affect the bias of parameter estimates and the width of credible intervals?

RQ2: At what PA:P ratio does 95% credible interval coverage drop below the nominal level?

RQ2a: At what PA:P ratio does parameter bias exceed a practically meaningful threshold?

> ***Hypothesis:*** *Increasing the PA:P ratio artificially inflates effective sample size, attenuating parameter estimates toward the null and producing misleadingly narrow credible intervals — a false sense of certainty driven by sample size, not ecological signal.*

Study Design

Data: 78 loggerhead nests + 78 false crawls from Santa Rosa Island and Perdido Key, Escambia County, FL (Cox, Schmutz, and Seals 2025)

. . .

Five LiDAR-derived predictors:

- Nest elevation (m above NAVD88)
- Dune height (m)
- Beach slope (degrees)
- Foreshore slope (degrees)
- Nest distance from MHHW line (m)

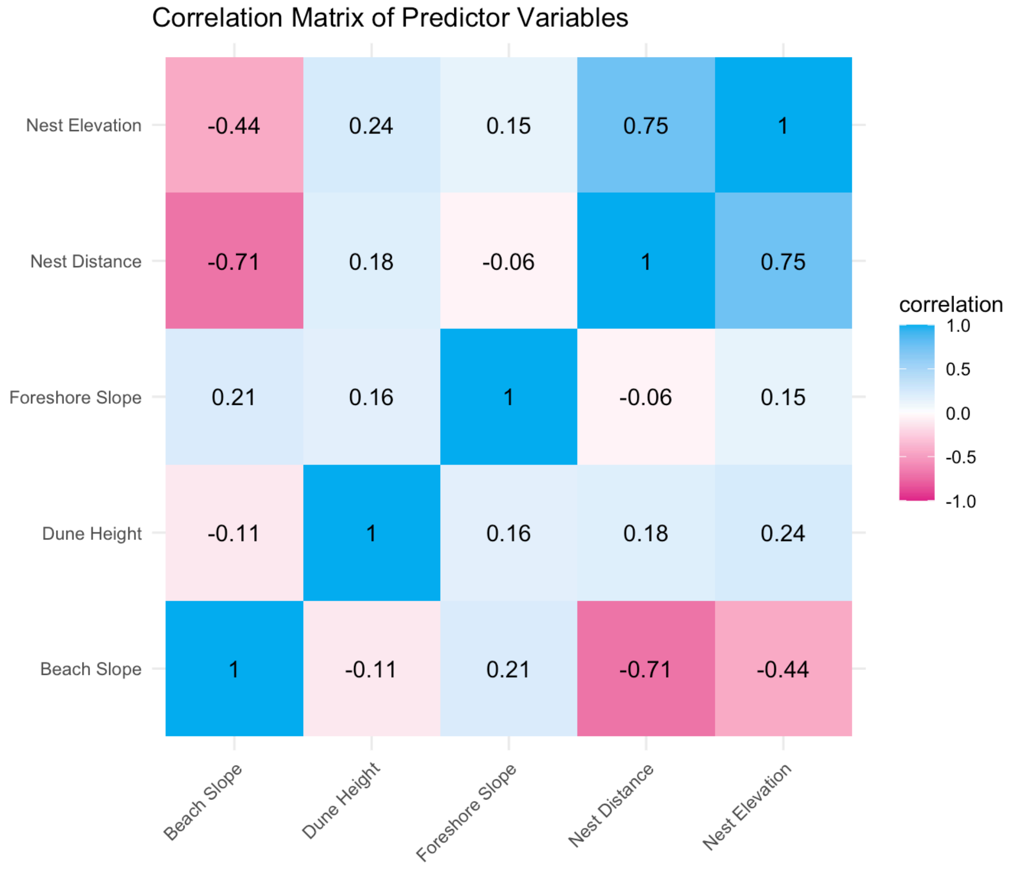
Dataset composition

PA:P	Nests	False Crawls	Total N	Prop.
1:1	78	78	156	50.0%
2:1	78	156	234	33.3%
5:1	78	390	468	16.7%
10:1	78	780	858	9.1%
20:1	78	1,560	1,638	4.5%

Model: Bayesian logistic regression via `rstanarm` with weakly informative Normal(0, 1) priors (Johnson, Ott, and Dogucu 2022; Northrup and Gerber 2018)

Simulation: 1,000 replicates per ratio. Metrics: bias, RMSE, power, and 95% CI coverage (Nemes et al. 2009; Puhr et al. 2017; Geroldinger et al. 2023).

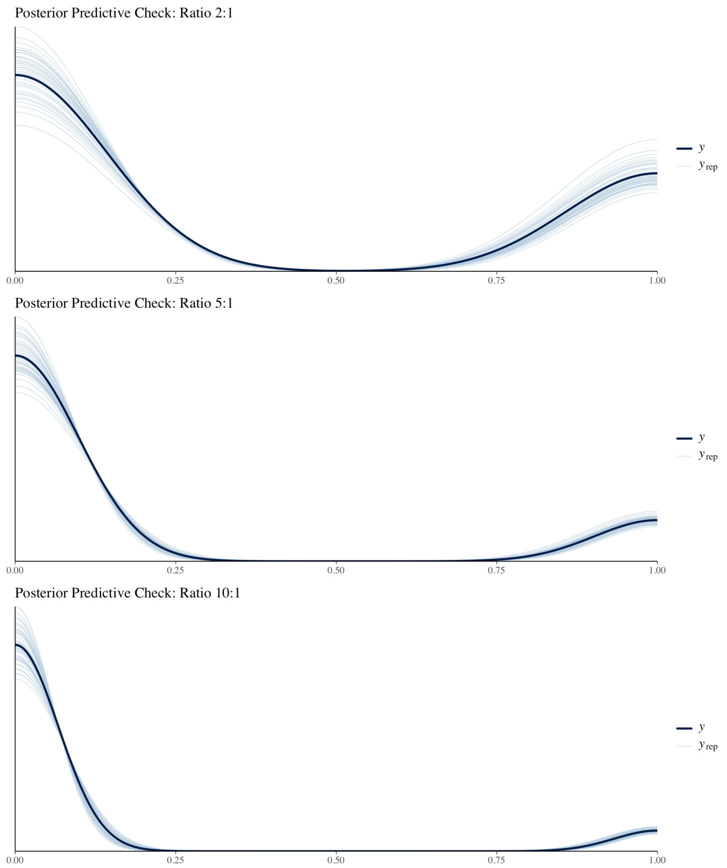
The Data: Weak Univariate Signals



Pairwise correlations among beach predictors. Nest distance and nest elevation are strongly correlated (r = 0.75); remaining predictors show weak associations (|r| < 0.30), confirming limited multicollinearity.

No single variable serves as a strong univariate discriminator — interquartile ranges of all five predictors **overlap substantially** between nesting outcomes.

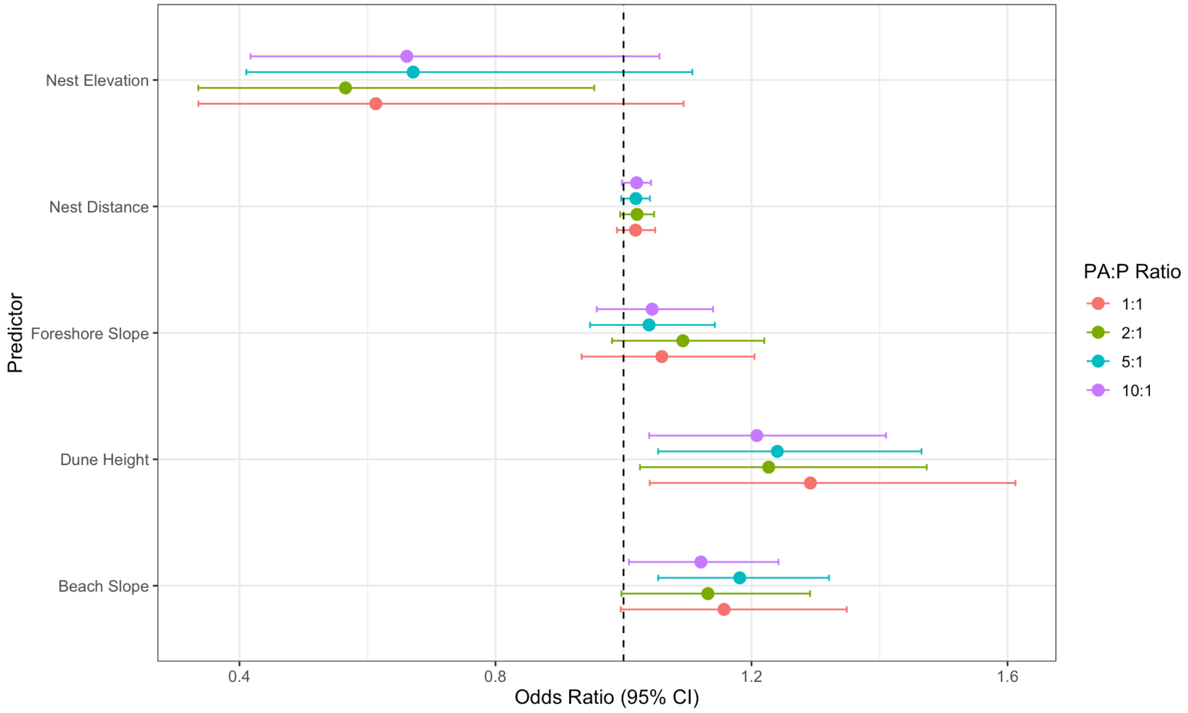
The Model Breaks Down as You Inflate



Posterior predictive checks: at 1:1, simulated data matches observed. At higher ratios, the model generates data concentrated near Y = 0 — confirming progressive misspecification driven by artificial class imbalance.

The model increasingly generates simulated data concentrated near Y = 0, reflecting the **artificial dominance of non-nesting observations**.

What Inflation Does to Your Estimates



Odds ratios with 95% CIs across PA:P ratios. Estimates attenuate toward OR = 1 while intervals narrow.

Attenuation: every predictor shifts toward the null

- Dune height: OR 1.29 ⟶ 1.21
- Nest elevation: OR 0.61 ⟶ 0.66

. . .

False precision: intervals narrow without new information

- Nest elevation CI width: 21% reduction
- Dune height CI width: 35% reduction

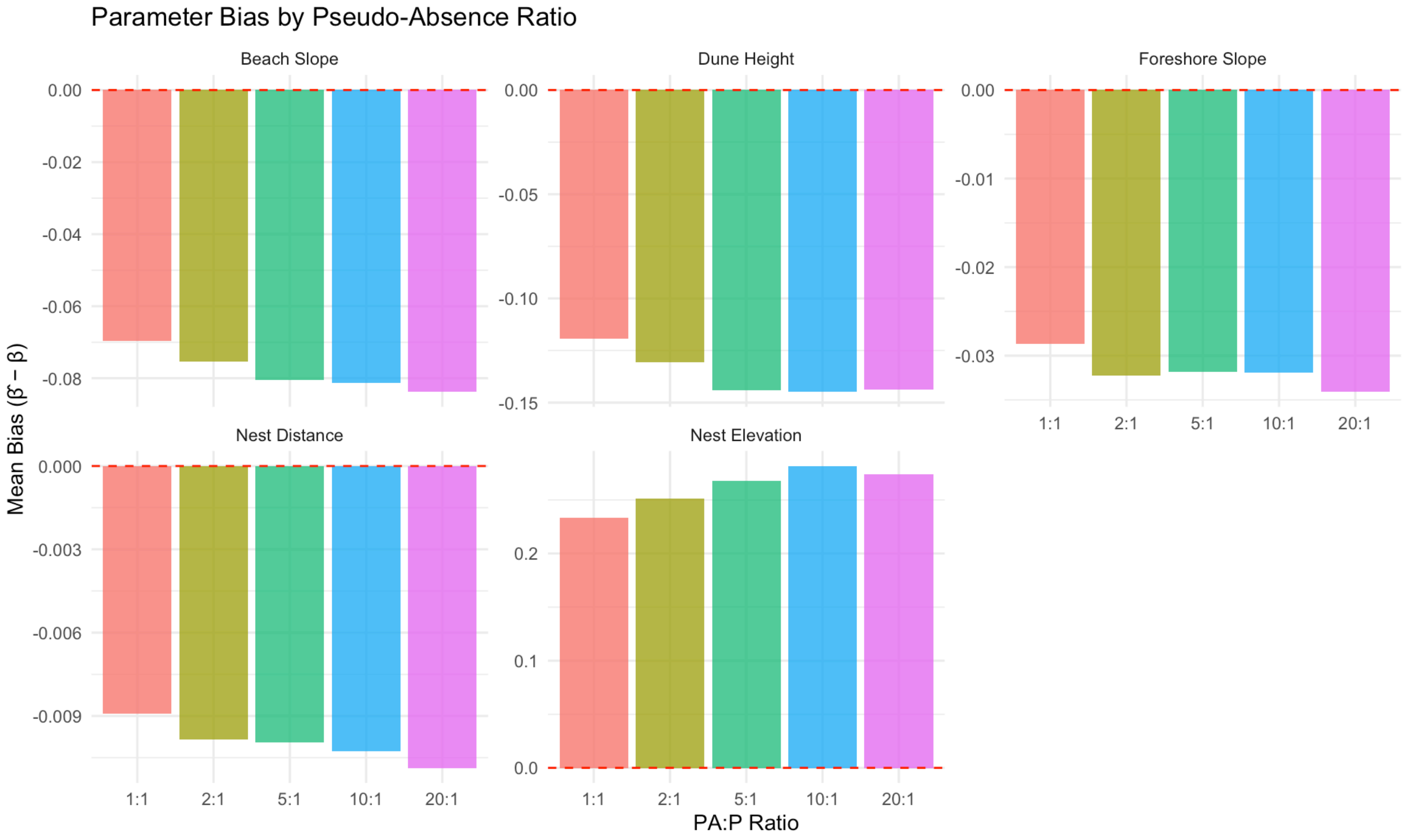
. . .

WAIC confirms it (Geroldinger et al. 2023):

Ratio	WAIC	SE
1:1	219.59	6.69
10:1	521.82	38.47

Each step of inflation worsens out-of-sample fit.

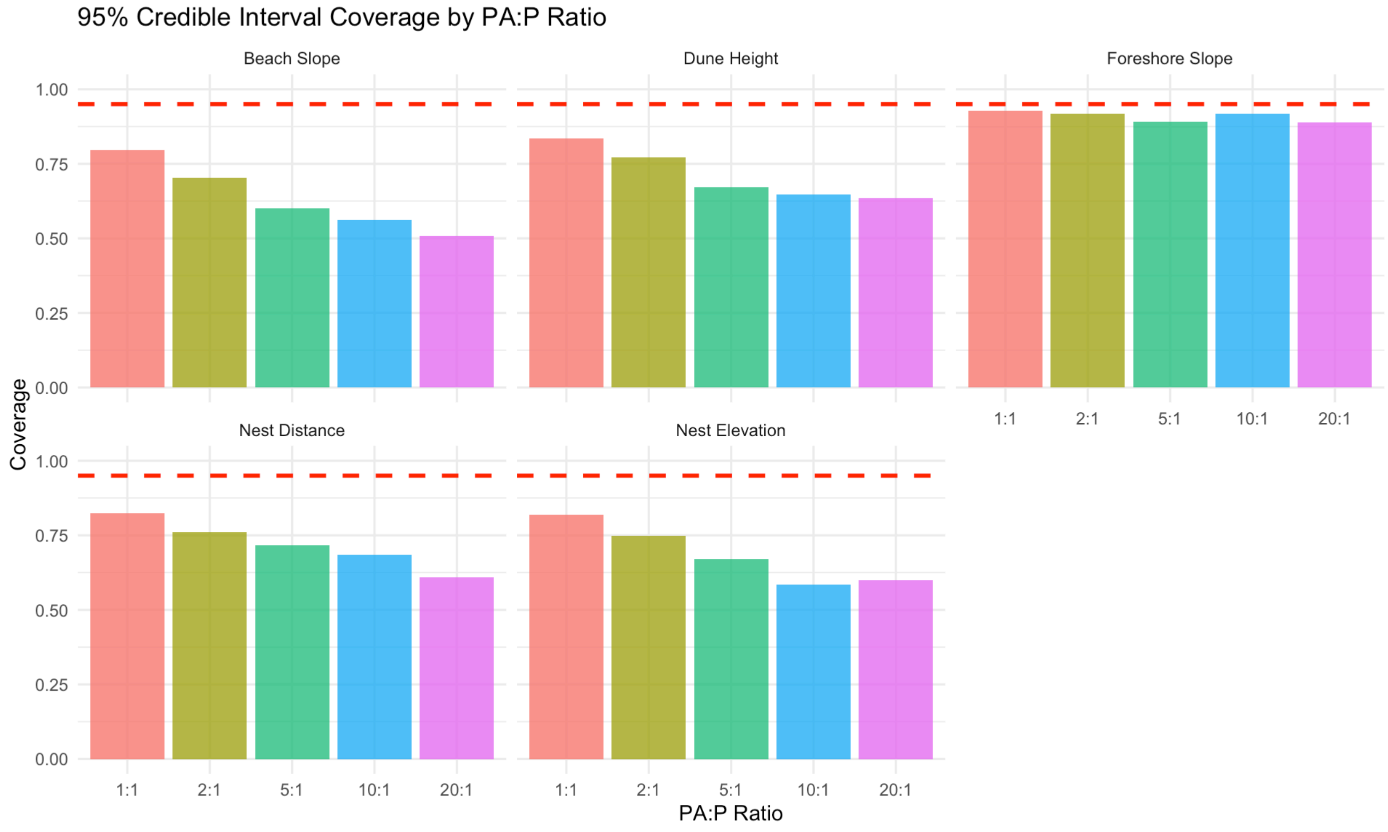
Simulation Evidence: Bias Is Systematic



Bias in parameter estimates increases monotonically with PA:P ratio across all five predictors. Attenuation toward zero is consistent with King and Zeng (2001) and Firth (1993).

Nest elevation (true β = –0.490): mean estimates recovered only **44–54% of the true effect** across conditions.

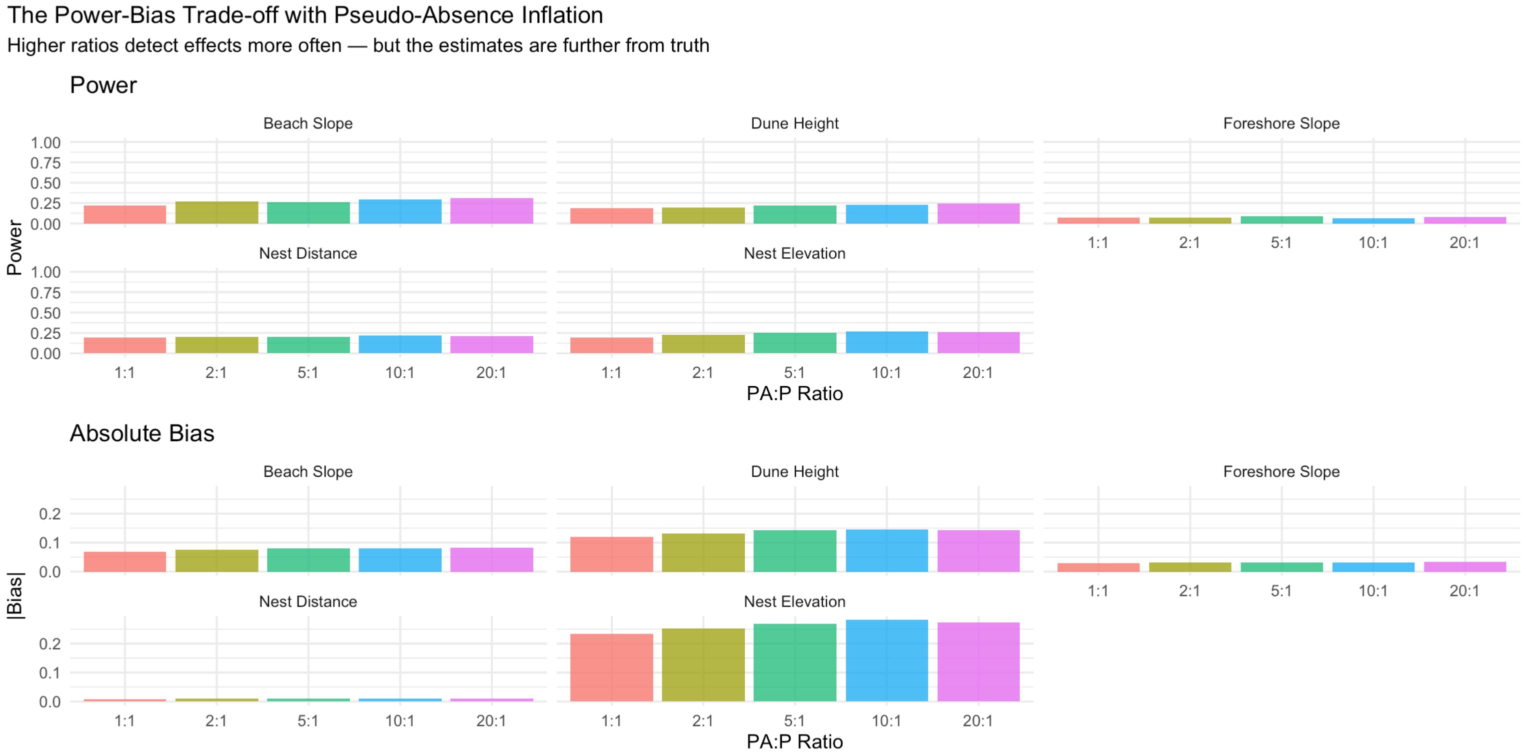
Your Credible Intervals Are Lying



95% CI coverage by PA:P ratio. The red dashed line marks the nominal 0.95 level. Coverage never reaches nominal at any ratio and deteriorates monotonically — consistent with Northrup and Gerber (2018) and Pavlou et al. (2016).

- Coverage **never achieved 95%** — not even at baseline 1:1
- Beach slope at 20:1: the "95% CI" contained the truth only **50.8% of the time**
- Primary predictors fell below 0.80 coverage by ratio **2:1**

The Power–Bias Trade-off



Power rises with PA:P ratio while bias grows in parallel.

Higher PA:P ratios **detect effects more often** — but the detected effect is further from the truth (Faghih et al. 2020).

A high-power, high-bias model generates **confident but incorrect conclusions**. A low-power, low-bias model generates **uncertain but honest ones**.

This dynamic could contribute to publication bias: journals favoring significant results combined with conventions favoring higher PA:P ratios would produce a literature that **systematically overstates certainty**.

So What? The Real-World Cost

Biased parameter estimates produce biased management decisions.

- Dune height attenuated from OR 1.29 to 1.21 — a manager using the 10:1 model would **underestimate the importance of dune restoration**
- Halls and Randall (2018) found beach nourishment positively associated with nest density, confirming that geomorphic effect magnitudes are **consequential for resource allocation**
- The "near-significant" results in Cox, Schmutz, and Seals (2025) (dune height $p = 0.066$, beach slope $p = 0.071$) may not reflect weak ecological effects — they may reflect **inferential artifacts of the PA:P ratio choice**

> **Important**
>
> More fabricated zeros are not better data. Treating resampled pseudo-absences as informative observations introduces inferential hazards that **standard model diagnostics do not detect**.

What Should Researchers Do?

For current practice:

1. **Report simulation-based bias and coverage** alongside standard SDM results — standard diagnostics miss this problem (Geroldinger et al. 2023)
2. **Use the observed PA:P ratio** as the more defensible inferential baseline
3. **Evaluate pseudo-absence plausibility** — the contamination problem means added points may represent true presences (Northrup et al. 2022)
4. **Separate predictive from inferential goals** — a design optimized for prediction may be the worst design for inference (Barbet-Massin et al. 2012)

For future work:

- Hierarchical extensions to pool information across multi-year datasets
- Spatially constrained PA strategies as alternatives to random resampling
- Extend this simulation framework to other SDM algorithms and ecological systems

Key Takeaways

Finding	Evidence
Bias is present at baseline	46–56% attenuation of true effect even at 1:1
Coverage never reaches 95%	Below 0.80 for primary predictors by ratio 2:1
Intervals narrow falsely	Up to 35% narrower — precision without accuracy
Power gains are misleading	Higher power detects a more distorted signal
Model fit degrades	WAIC: 219.59 → 521.82
Conservation impact is real	Underestimates dune importance for nest protection

The pseudo-absence ratio is not a tuning parameter. It is an inferential assumption that demands justification.

References

Barbet-Massin, Morgane, Frédéric Jiguet, Cécile H Albert, and Wilfried Thuiller. 2012. "Selecting Pseudo-Absences for Species Distribution Models: How, Where and How Many?" *Methods in Ecology and Evolution* 3: 327–38. https://doi.org/10.1111/j.2041-210X.2011.00172.x.

Cox, Darcy, Phillip Schmutz, and Samantha Seals. 2025. "Modeling Sea Turtle Nesting Probability: Evaluating Loggerhead Nest Site Selection Using Presence-Pseudo-Absence Data." *Authorea Preprint*. https://doi.org/10.22541/au.175525924.40059675/v1.

Elith, Jane, and John R Leathwick. 2009. "Species Distribution Models: Ecological Explanation and Prediction Across Space and Time." *Annual Review of Ecology, Evolution, and Systematics* 40: 677–97.

Faghih, Ali, Zahra Bagheri, Ana Stevanovic, Seyyed Mohammad Taghi Ayatollahi, and Peyman Jafari. 2020. "A Comparative Study of the Bias Correction Methods for Differential Item Functioning Analysis in Logistic Regression with Rare Events Data." *BioMed Research International* 2020: 1632350. https://doi.org/10.1155/2020/1632350.

Firth, David. 1993. "Bias Reduction of Maximum Likelihood Estimates." *Biometrika* 80 (1): 27–38. https://doi.org/10.1093/biomet/80.1.27.

Gelman, Andrew, John B Carlin, Hal S Stern, and Donald B Rubin. 2004. *Bayesian Data Analysis*. 2nd ed. Chapman; Hall/CRC.

Geroldinger, Angelika, Lara Lusa, Mariana Nold, and Georg Heinze. 2023. "Leave-One-Out Cross-Validation, Penalization, and Differential Bias of Some Prediction Model Performance Measures: A Simulation Study." *Diagnostic and Prognostic Research* 7: 9. https://doi.org/10.1186/s41512-023-00146-0.

Guisan, Antoine, and Niklaus E Zimmermann. 2000. "Predictive Habitat Distribution Models in Ecology." *Ecological Modelling* 135: 147–86.

Halls, Joanne N, and Matthew Randall. 2018. "Modeling Sea Turtle Nesting: Temporal and Spatial Patterns Along the North Carolina Coast." *ISPRS International Journal of Geo-Information* 7 (9): 348. https://doi.org/10.3390/ijgi7090348.

Johnson, Alicia A, Miles Q Ott, and Mine Dogucu. 2022. *Bayes Rules! An Introduction to Applied Bayesian Modeling*. CRC Press.

King, Gary, and Langche Zeng. 2001. "Logistic Regression in Rare Events Data." *Political Analysis* 9 (2): 137–63.

Nemes, Szilárd, Jón Magnusson Jonasson, Anna Genell, and Gunnar Steineck. 2009. "Bias in Odds Ratios by Logistic Regression Modelling and Sample Size." *BMC Medical Research Methodology* 9: 56. https://doi.org/10.1186/1471-2288-9-56.

Northrup, Joseph M, and Brian D Gerber. 2018. "A Comment on Priors for Bayesian Occupancy Models." *PLOS ONE* 13 (2): e0192819. https://doi.org/10.1371/journal.pone.0192819.

Northrup, Joseph M, Eric Vander Wal, Maegwin Bonar, John Fieberg, Michel P Laforge, Martin Leclerc, Christina M Prokopenko, and Brian D Gerber. 2022. "Conceptual and Methodological Advances in Habitat-Selection Modeling: Guidelines for Ecology and Evolution." *Ecological Applications* 32 (1): e02470. https://doi.org/10.1002/eap.2470.

Pavlou, Menelaos, Gareth Ambler, Shaun R Seaman, Oliver Guttmann, Perry Elliott, Michael King, and Rumana Z Omar. 2016. "How to Develop a More Accurate Risk Prediction Model When There Are Few Events." *BMJ* 351: h3868. https://doi.org/10.1136/bmj.h3868.

Puhr, Rainer, Georg Heinze, Mariana Nold, Lara Lusa, and Angelika Geroldinger. 2017. "Firth's Logistic Regression with Rare Events: Accurate Effect Estimates and Predictions?" *Statistics in Medicine* 36 (14): 2302–17.