Dear Dr. Nicolas Gotelli,

I hope this finds you well.

Thank you and the reviewers again very much for the constructive feedback on our manuscript and the opportunity to revise and resubmit it to *Ecology*.

In general, we agreed with the points raised by the reviewers, and typically included the changes requested. If not possible (e.g., if requests were beyond the scope of the present manuscript), we included text in the manuscript to address the issue and discussed the situation in our responses detailed below. We believe that the changes to the manuscript have increased its readability, strengthened its technical foundation, and further supported its general applicability to ecological forecasting.

I confirm that this manuscript has not been published elsewhere and is not under consideration by another journal. All three authors have approved the revised manuscript and agree with its resubmission to *Ecology*. On the behalf of my co-authors, I thank you for the opportunity to resubmit this manuscript and look forward to hearing from you.

Juniper Simonis, on behalf of Ethan White and Morgan Ernest

Reviewer 1:

Introduction

Provide an overall step-by-step outline

**We agree that the manuscript would benefit from having a set of explicit steps outlined, and we have edited it to address on line XXXX.**

Cross validation options

The reviewer stated that we did not discuss cross-validation options but instead referenced other papers.

**Indeed, given that this is a report-length manuscript, we do not have significant length to discuss in significant detail all aspects, but rather must cite papers in reference [in contrast to what the reviewer states, we *do* discuss cross validation options (see lines XXXX)].**

**Considering that this paper follows standard practice in fields with extensive forecasting history, however, we focus on end-sample hold out methods that have been shown to produce more realistic forecasts than cross-validation approaches (as cited on lines XXXX).**

**Note that the Cerequeria et al. 2019 preprint the reviewer cites aligns with what we proposed herein (the reviewer states the opposite), as evidenced by the end of their abstract: “the most accurate estimates are produced by out-of-sample methods”**

**The Roberts et al 2017 review is an important and welcomed contribution to cross-validation, but we caution the reviewer on their transfer of cross-validation concerns directly to end-sample approaches, which have noted (lines XXXX) different philosophical and statistical considerations.**

**Indeed, while we appreciate the reviewer’s work and references provided around cross-validation, the prevailing practice in well-established forecasting fields is end sample hold-out, and we believe that a table of cross-validation approaches would be a distraction in a paper that does not use such approaches.**

**We do, however, acknowledge and appreciate the reviewer’s desire for more discussion of this topic, and so we have detailed more of the issues around using cross-validation in forecasting and the new work being done to address it on lines XXXX.**

Residuals

The reviewer would like us to describe the model residuals in evaluation.

**We agree this is a critical part of evaluation, and apologize for omitting it previously. We have added text to address this on line XXXX.**

Clarify generation of H

The reviewer was confused about how the prediction distribution H was generated.

**We have added text on lines XXXX to address this confusion and highlight the generality of the terminology used.**

Clarify application of PIT

The reviewer was confused about how the PIT applied to or was associated with scoring rules.

**We apologize for any confusion around this, but have added text on line XXXX to clarify that the PIT is not applied to scores but to the predicted distribution.**

Discussion of the difference between within and out of sample evaluations

The reviewer stated their belief that the difference between the two PITs and scores for each of the sets shown in Figure 2 were due to bias introduced by the analyst, and expressed concern regarding stationarity.

**An important point of note is that the PITs are all out-of-sample, each generated using held out end-of-sample data. No evaluation metrics are produced from “within sample” (training data).**

**Further, we do not agree that any forecast assumes stationarity, even in a wider sense. Many models, even simple ones, accomplish non-stationarity by allowing changes in parameters (at necessary levels, given the system/model) and including non-reversible changes.**

**Thus, while we do agree that stationarity requires detailing (text added on lines XXXX), we do not agree with the reviewer that the mention should be one to indicate that forecast models are stationary, because that is not the case.**

Abstract

The reviewer requested we mention more specifics of methods and recommendations in the abstract.

**We have added text on line XXXX to address this.**

Introduction

The reviewer requested we define “forecast”.

**We have added text on line XXXX to address this.**

The reviewer requested we add “as found by” into the brackets on line 30

**We have added text on line XXXX to address this.**

Context and Terminology

The reviewer requested we clarify the distribution terminology for G on line 43.

**We have added text on line XXXX to address this.**

The reviewer noted repetition at the end of line 47.

**We edited the text on line XXXX to address this.**

The reviewer asked for clarification on what ‘both’ meant on line 48.

**We edited the text on line XXXX to address this.**

The reviewer questioned why we singled out the prequential approach.

**We added text on line XXXX to explain here (although it is said in the subsequent section as well) that this approach is the well-established and evaluated method in fields with far more extensive backgrounds in forecasting.**

“Buffer”

In multiple spots, the reviewer states that the “buffer” from the forecast origin to the first predicted sample does not need to be 1 step.

**However, this appears to be from a confusion around regarding how forecasts are made and evaluated. Given the time series nature, one cannot skip a datum and, for example forecast o + 2 without first forecasting o + 1. Not all forecasted steps need to be evaluated (and indeed, the reviewer is right that some may wish to not, which we have added text to address on line XXXX), but all steps from the origin to the evaluated predictions need, fundamentally, to be forecast through.**

Forecast Validation

The reviewer requested that we do not call the prequential approach the dominant approach, and play up the usage of cross-validation approaches.

**We disagree with the reviewer’s statement and request here--- as we have cited and stated, it is correct to say that the dominant tool used in forecasting science (broadly speaking, which includes many fields and only recent forays by ecology) is end-sample-hold-out, which has been repeatedly evaluated and show better performance than cross-validation approaches.**

**We do, however, understand the need to develop this a bit more directly so that readers understand the breadth and depth of knowledge on forecasting tools that resides outside of our field. We have edited lines XXXX to address this.**

The reviewer was confused about why the length of the training data and testing data need to have particular lengths in relation to each other.

**We have added text on line XXXX to address this.**

The reviewer did not understand why an analyst would not update a model.

**Not all forecasting models are as easy and as simple to update as the reviewer paints them. Indeed, quite a few forecasting models used rely on, for example climate dynamic models that are exceptionally large and take far longer to update than to evaluate (and would not show significant change after updating).**

**We have added text on line XXXX to explain this and contextualize where and when that approach might be needed**

“Cyclic Dynamics”

The reviewer stated in a few places that they did not agree with our use of the word cyclic.

They believe that cyclic dynamics are not common in ecology

They believe that the data shown in the example are “seasonal not cyclic”

**Cyclic dynamics occur on many time scales, and we believe that the reviewer is placing an uncommon and unnecessary specification of “interannual” to our term “cyclic”, which we did not add and are not specifically discussing.**

The reviewer was confused about why cyclic dynamics would make a forecast origin be important (as stated on line 84).

**As stated at the beginning of that paragraph, this is discussion of a single end-sample hold out, for which the origin will play an outsized role, especially with dynamical systems, as where the system is on a trajectory will bias how well the next forecast matches. We have added text on line XXXX to address this.**

Graphical Evaluation

The reviewer was confused why we mentioned plotting the training data, saying that if the model fits well to the past data that it’s probably poor and that ideally it would be uninformative.

**We disagree with the reviewer here and believe they misread what we stated in the manuscript, which was not to see the evaluation of fit to training data but “to show past dynamics”. Indeed, it is critical for a forecaster to understand if the model is able to capture the general range of count and dynamical variation. We made no statement about fit in the reasoning for showing past dynamics.**

The reviewer asked us to rephrase our description of how the predicted-vs-observed plot should ideally look .

**We have edited the text to address on line XXXX.**

The reviewer asked us to add reference to Gelman’s work on posterior p values to our discussion of the PIT

**We have added text to address on line XXXX.**

Scoring Rules

The reviewer asked us to clarify terminology used in explaining the Brier Score

**We have edited the text to address on line XXXX.**

The reviewer was confused about what variance was “inflated”. The reviewer also seemed to believe that we were saying that the altering of variance was bad.

**We have added text to address the confusion around variance inflation on line XXXX. Note too, however, that we did not say that it was something faulty that there was altered variance. Rather, we merely said that accounting for it requires more samples. This is a straightforward statement of fact.**

Comparing Model Scores

The reviewer asked for clarification on the phrase “marginal predictive distribution” rather than just a citation.

**We have added text to address on line XXXX.**

The reviewer was confused by our general use of the phrase “correlation”, and was wondering if that meant “temporal autocorrelation”.

**Indeed, temporal autocorrelation is one of the multiple ways in which samples can be non-independent that needs to be included in a model. There are other sources of correlation as well (spatial, phylogenetic, sampling based, etc), and it is not appropriate to single out temporal autocorrelation in such a way as suggested by the reviewer. Rather, we highlight that it is one of many sources on line XXXX.**

The reviewer did not understand the purpose of the D-M test being discussed as is if it needs robust formulae.

**Note, however, that the opening to the paragraph on the D-M test explicitly says that the test is good as long as correlations are accounted for. And indeed, this same stipulation is made throughout the D-M paragraph. Despite the reviewer’s statements, the paragraph is consistent in this regard: the D-M test can compare forecasts as long as correlations are accounted for and there are methods to do that. Obviously the “formal” test stat is the simple version of the model with no correlation and just the standardized difference, and the robust formula generalize it.**

Example

The reviewer was confused about the application and interpretation of the PIT.

**Despite the reviewer’s statement, the PIT does itself show bias. This is based on the standard (cited) uses of the PIT (as detailed within the manuscript) and its evaluations in forecasting, and is well displayed in Figure A1. As shown in Figure A1 as well, an imprecise (too wide coverage) prediction would result in a central peak to the PIT, not a peak at the upper range, which is shown here, indicating negative bias.**

**Again, there is no “in-sample” or “out-of-sample” PIT. All of the PITs are from out of sample comparisons. The difference between the two is, as stated in the text, due to the vagaries of the particular final hold out evaluation, which as stated, is why an measure of the model broadly should involve many evaluations.**

**To facilitate proper interpretation of the PIT of the example, we have included references to Figure A1 in this section.**

Discussion

The reviewer stated that our discussion was “boring” and that they “would not mind to see it go”.

**These are both non-constructive feedback and preferably would be avoided. While we appreciate the reviewer’s point that they would have preferred we cover topics differently in the discussion, there are far more constructive and supportive ways to provide actionable feedback. Further, we note that with any paper there are many topics that could be discussed (and prerogatives for each), and there is never room for all of them. We acknowledge the reviewer’s point here but do not take kindly to their tone or approach.**

The reviewer requested that we offer recommendations about the steps they outlined early in their review.

**We agree that the discussion would benefit from actionable steps and have added text to that effect on lines XXXX.**

Tables

The reviewer was confused about the notation for the quadratic score and requested clarification.

**It is not entirely clear what the reviewer is confused with in Table 1. But no, f(y\_n) is not a vector. y\_n is (as stated) “an observed value” and f() is (as stated) “probability density or mass function”. Thus, f(y\_n) is a function observed at a value (one singular value), so f(y\_n) is a scalar. The reviewer then states that ||f(y\_n)|| sums over y, referencing the footnotes, but that is not at all stated in the footnotes. Rather, all the footnotes show is the calculate of a p-norm, which sums over the p, not over the y. summation over y is not shown in the able or footnotes.**

Figures

The reviewer asked for some clarification on Figure 2:

-adding labels and values to the axes in b and c

-expanding panel d to show the different models better

-align background coloring of b and c with a and d

**We addressed these in edits**

The reviewer incorrectly states that we could add “monthly” in the legend for abundances.

**However those values are not monthly, rather lunar-monthly. As such, given the unnecessary confusion around that nuance, the addition of extra terms to the legends is not necessary and would be distracting, so we have not done so.**

Minor

The reviewer was confused about our use of the word “recurrent” on line 94.

**We are using the word recurrent as its dictionary definition to mean recurring/re-occurring. And it is clearly stated in the sentence (as is) that the forecasts are what are recurrent. Recurrent does not mean updating, it means occurring again and again, rather than all at once (as stated in the manuscript)**

**Thank you for catching the error on line 123, it has been edited.**

**Thank you for catching the use of “differential” when “difference” would be more appropriate.**

**Thank you for noting the need to specify right truncation of the Poisson.**

Reviewer 2:

Abstract

Line 4: edit to focus on literature more

**Agreed that the review needs focus here; edited to address on line XXXX.**

Introduction

Line 29: edit to remove redundancy

**Agreed that language can be tightened a bit to avoid repetition; edited to address on line XXXX.**

Lines 36-38: describe the outline of the ms and align with section headers.

**Agreed that the end of the introduction needs more direction; edited to address on line XXXX.**

Scoring Rules

Line 121: lead the explanation of scoring rules with concept, not terminology

**Understood how that is overwhelming; edited to address on line XXXX.**

Figures

Figure 1: Panel (a) is distracting and does not add much

**Definitely understand that; panel (a) has been removed to give more space and focus to other panels.**

Figure 2: use letters to specify the additional panels rather than squeezing them into (b) and (c).

**We understand the reviewer’s point here and have added panel lettering to address.**

Pocket Mouse Example

Add more text from Appendix B to main text to distinguish differences with White et al. 2019 and include details on sampling, etc.

**Agreed that the example needed more treatment in the main text, we have added details to address these two issues on line XXXX.**

Discussion

Expand to include how case study does or does not represent forecasting in ecology in general

**We agree that this is an important point to highlight and have added text to address on line XXXX.**

Expand to discuss next steps/future work

**We are in agreement with the reviewer and to facilitate this, we have added text on line XXXX.**

General

Intro or Discussion: mention observation error more explicitly

**Definitely agree with the important of observation uncertainty, we have added details to discuss this topic on line XXXX.**

Minor

**Thank you for catching the error on line 123, it has been edited.**

And thank you, reviewer 2 for the complements on the workflow!

**on line XXXX**