Dear Dr. Corley,

Please find attached our revised manuscript, “Increased adoption of best practices in ecological forecasting enables comparisons of forecastability.” In response to the Reviewers’ comments, we have made many changes to the manuscript, which we think have substantially improved the text. Importantly, we have published the data associated with the study to the Ecological Data Initiative (EDI) repository and submitted a new supplemental file with the matrix screening questions used in this study, resolving Reviewer 2’s concern about data accessibility.

In response to Reviewer 1’s question about how our best practices were selected, we created a new supplemental table that outlines all practices identified in the four papers we are aware of that provide best practices for near-term ecological forecasting. We added one new best practice (“Compare modeling approaches”) to our manuscript and removed one practice (“Partition uncertainty”), so that we now analyze all practices that are mentioned in at least two of the four papers. “Compare modeling approaches” is demonstrated in 47% of papers and exhibits a non-significant upward trend over time. Adding this practice did not change any conclusions of the paper, but makes our selection of proposed best practices more robust. Likewise, removing “partition uncertainty” did not alter our conclusions. Below, we respond to each Reviewer’s point in bold. We hope that the revised manuscript is now suitable for publication in *Ecological Applications* and look forward to hearing from you.

Sincerely,

Abby Lewis, on behalf of the co-authors

Reviewer(s)' Comments to Author:

Reviewer: 1

Comments to the Author

This paper describes a literature search, both structured and secondary based on the first search, to detect how often a set of best practices are used for near-term ecological forecasting. It is clearly written, well described, and the choice of analytical metrics as well defended. I suggest only minor revision, including some expansion of justification of the selection criteria of the nine best practices selected.

**We appreciate the Reviewer’s positive feedback! We have made multiple changes throughout the manuscript to expand our justification of the best practices included in this analysis (described below, with line numbers provided).**

The title is the slight overreach. This is because the comparisons are forecasting across systems is relatively limited in the paper. This is more an aspirational claim than a delivered one. I’m not sure that forecastability is a word we need either!

**We appreciate the Reviewer’s comment, and have removed “across systems” from the title. We consider the analysis of forecastability across scales and variables to be an important contribution of this paper, and have therefore retained the word “forecastability” in the title. To address the Reviewer’s concern, we have clarified how we are defining forecastability in the Abstract and Introduction of the manuscript, saying that it is “defined here as realized forecast accuracy” (lines 39, 94–95). We have also added additional text to the Introduction to clarify the relationship between forecastability and predictability. The modified text now reads: “Ecological forecasting can be a particularly powerful test of predictability, as forecasting requires predicting beyond the range of observed data. Thus, comparisons of forecastability complement and extend existing theoretical and modeling-based work that has predicted how various factors (e.g., forecast horizon, computational irreducibility) influence the relative predictability of ecological variables (Beckage et al. 2011, Petchey et al. 2015)” (lines 99–103).**

**With respect to whether forecastability is a useful word more generally, we have chosen to use this word for two reasons: first, ecological variables may be highly predictable based on existing data without being highly forecastable (e.g., if they are sensitive to an environmental driver that cannot be forecasted with sufficient accuracy), and second, use of the word “forecastable” is already common in previous published literature (e.g., see Clark et al. 2001, Petchey et al. 2015).**

The abstract is clearly written. The near-term definition could be introduced here too. Later in the paper, near term is defined as less than 10 years, which in my view conflates two timescales seasonal and decadal. Much of the analysis is then also completed on forecasts with a timescale of less than seven days, which is weather time scale forecasting.

**We thank the Reviewer for their positive feedback. To address the Reviewer’s comments, we now specify in the Abstract that we are defining near-term as forecasts with a horizon ≤10 years (line 26). This is in agreement with previous literature, particularly Dietze et al. (2018), who define near-term as “daily to decadal.” While we agree that this may include weather-scale, seasonal, and decadal patterns, the timescale that constitutes a near-term forecast likely depends upon the scale of the process being measured. For example, a near-term forecast of lake chlorophyll-a would typically be seasonal or subseasonal, while a near-term forecast of tree diameter might be annual or longer. Consequently, using an expansive criterion (10 years) allowed us to encompass most near-term forecasts on a relevant timescale for each forecast variable. Furthermore, the majority of papers (75%) in our final dataset predicted within one year into the future, suggesting that a larger body of decadal and multi-decadal forecasting literature was effectively screened out in our literature search.**

I think the paper would be improved if it had a table, or box of interesting forecasting examples that you came across in your search. You could profile three or four case studies.

**While we thank the Reviewer for this suggestion, we feel uncomfortable choosing just 3 of 178 near-term ecological forecasting papers to highlight, as we want to focus our manuscript on the data-driven analysis, rather than a qualitative review of near-term ecological forecasting literature. That being said, we have now published the entire dataset of papers to the Environmental Data Initiative repository for readers to explore (see Lewis et al. 2021). Additionally, we added citations for three papers to the Results section to highlight how they quantified uncertainty (lines 340–344), which we hope helps direct readers to papers that might be of interest for demonstrating specific best practices.**

**Lewis, A. S. L., W. M. Woelmer, H. L. Wander, D. W. Howard, J. W. Smith, R. P. McClure, M. E. Lofton, N. W. Hammond, R. S. Corrigan, R. Q. Thomas, and C. C. Carey. 2021. Systematic review of near-term ecological forecasting literature published between 1932 and 2020. Environmental Data Initiative. https://doi.org/10.6073/pasta/c4bea94f100f39a6b73c7b9a577df214**

The selection of the best practices is not clear. While the box 1 indicates which papers identify particular features, you have not chosen the set that is used in all these papers. How did you arrive at the set that you propose as best practice? I think you need to have more explanation on how you arrived at the set.

**We thank the Reviewer for this very helpful comment, which inspired us to create a table summarizing and synthesizing the best practices presented in Harris et al. (2018), White et al. (2019), Hobday et al. (2019) and Carey et al. (2021). This table is now included in a new supplement, Appendix S1. We reference this new table in the manuscript main text on lines 80, 217, and 554.**

**In creating this table, we found that we had included all best practices that were referenced in at least two papers except “compare multiple modeling approaches.” We had already assessed whether multiple approaches were compared in each paper using our matrix analysis. Therefore, we were able to add this as a best practice in the manuscript. In the revised manuscript, “Compare modeling approaches” has been added as a best practice in Box 1 (lines 856–861), and how we assessed whether papers demonstrated this practice has been added to the Methods (lines 243–244). We report the percentage of papers that demonstrate this practice on line 363 and the number of approaches they compare on lines 372–374. Summary statistics for logistic regression results have been added to Table 1 (line 866), and a brief discussion of these results is added on lines 432–436.**

**One of our previous best practices was only included in one of the four best practice sources: “partition uncertainty.” Following our new criteria of including all practices that were referenced in at least two papers, we have since removed this best practice. In the updated manuscript, we retain previous text relating to uncertainty partitioning in the Results and Discussion. However, we now refer to uncertainty partitioning as an additional step that expands on “include uncertainty” rather than as its own best practice (lines 339–344, 473).**

Just a comment on both forecast requirement number one, uncertainty, and number two, assess and report forecast skill. From the perspective of end users scale is sometimes all they are interested in. The measures of skill will incorporate the uncertainty in the overall forecast. The uncertainty, and I agree with your suggestion of breaking down uncertainty, will be most useful to forecast developers when they try and improve the forecast. I think you could expand on this point in the discussion.

**We appreciate the Reviewer’s comments about how uncertainty can be useful for forecast improvement, and have added text on this point to the Discussion. The section now reads: “Moving beyond specifying uncertainty to partitioning uncertainty into its component parts (e.g., initial condition, driver, parameter, and process uncertainty) provides information to help forecast developers prioritize improvements to their forecasting system and allows researchers to understand the constraints to predictability for a given ecological variable (Dietze 2017b)” (lines 473–477).**

You might also identify best practices that are good for end users, and best practices which serve the research community. Do you have a mix of these in your set. You would be able to add some material in the discussion noting that the different groups would require different best practices.

**We agree wholeheartedly with the Reviewer that some best practices are more useful for end users and some are more useful for research purposes. For this reason, we had previously divided the best practices into three categories (Forecast Requirements, Decision Support, and Research; Box 1). To address the Reviewer’s comment, we have also added additional text on this point to the Discussion to make this delineation more clear. The modified text now reads: “While the rates of adoption of these proposed best practices (Box 1) are low overall, they are not necessarily unexpected. Different forecasting applications likely require different best practices; in this analysis, we have divided our selected best practices among three categories: forecast requirements, decision support, and research. However, this is a coarse delineation, and the last two tiers are not mutually exclusive: decision support practices can also be important for ecological understanding and vice versa” (lines 545–550).**

You have also minimised the ethical aspects of best practice forecasting in this paper.

**We strongly agree that ethical aspects of forecasting best practices are critically important, and our analysis reveals that they are underrepresented in current literature. In particular, we highlight that only 5% of papers overall and 25% of forecasts that are in use by an end user mention any ethical considerations (line 494). To respond to the Reviewer’s comment, we added text to be more explicit about how training is needed to advance the consideration of ethics in forecasting. This section of the Discussion now reads: “Given the power of forecasts to inform decision support, education on how to navigate engagement with end users, and particularly any ethical considerations that must be made, may be useful in improving the utility of forecasts for stakeholder use” (lines 495–497).**

The supplementary data are all provided, which allows good use into the future and reproducibility of these results.

**We thank the Reviewer for their positive feedback!**

Specific comments

Line 66. I consider Payne et al a good review of the state of marine ecological forecasting - maybe 4 years is a long time....

**We agree with the Reviewer that Payne et al. (2017) is an excellent review of marine ecological forecasting literature. We have chosen not to cite Payne et al. (2017) for this particular sentence because that analysis is neither systematic (the authors reviewed papers they were familiar with and conducted a brief literature survey), nor a comprehensive review of all ecological forecasting (it focuses on marine systems only). Instead, we have added this citation later in the Introduction (lines 84–85) to demonstrate how previous forecasting efforts can provide insight into the development of the field.**

Line 95. Can you suggest some classical ecology papers from the 1950s that defined this aspect of predictability for the field of ecology?

**There are many possible examples of classic ecological concepts that are centered in ecological predictability, including Lotka Volterra predator prey dynamics, island biogeography, and more. Here, we chose to add Clements (1936) and Gleason (1926), highlighting how the question of whether plant communities are primarily characterized by climax communities (Clements 1936) or individualistic responses (Gleason 1926) is fundamentally a question of ecological stability and predictability. The modified sentence reads as follows: “Understanding ecological predictability is a fundamental goal in ecology (Gleason 1926, Clements 1936, Sutherland et al. 2013, Godfray and May 2014, Houlahan et al. 2017, and references therein) and provides valuable information regarding the nature of ecological processes (Petchey et al. 2015)” (lines 95–98).**

Line 138. The definition of near term is less than 10 years. This is quite long and covers weather scale forecasting, seasonal forecasting, multi year, and decade. Each of these timescales have been a focus in their own right, as they require different focus on initial conditions, relative to boundary conditions.

**We hope the Reviewer supports our justification for the decadal threshold, which is detailed in a comment above. In brief, we are using this near-term threshold so our work aligns with previous ecological forecasting literature and is inclusive of the relevant timescale of each forecasted variable (e.g., phytoplankton chlorophyll to tree diameter). To respond to this point, we have also cited Dietze et al. (2018) here (line 140).**

Line 152. Just be clear that this was 21% of 142 papers rather than 21% of 2711 papers.

**We appreciate the Reviewer’s comment and have added “(n = 669)” on line 154 to clarify this point.**

Line 598. This is the claim made in the title that this attention to best practice will provide critical insight into the predictability. This aspect is underdone in the paper, and can either be removed or expanded. For example the variables you tracked over seven days Fig 7 were both marine and terrestrial. Does the difference in slopes indicate that the terrestrial environment is more predictable than the marine? Or are you analysing to few variables to make any suggestion (I think too few variables).

**We agree with the Reviewer that we are analyzing too few variables to make any comparison between aquatic and terrestrial ecosystems, which was not the intent of the manuscript. Consequently, we have modified the title and section headers so they no longer include mention of comparing forecastability “across systems.” Further, we have modified this concluding sentence to say that near-term ecological forecasting can provide insight into the predictability of ecological “variables” rather than ecological “systems” (line 619). While this is a preliminary analysis, we feel justified in saying that “near-term ecological forecasting is well-positioned to transform ecological management and provide critical insight into the predictability of ecological variables'' because (a) the rate of forecast publication is increasing, (b) adoption of best practices that facilitate comparisons across studies are increasing, and (c) forecast accuracy can be compared across scales and variables, as demonstrated in this analysis.**

Most figures are clearly drawn and high-quality.

**We appreciate the Reviewer’s positive feedback.**

Figure 1, a B and C could be separated a bit more in space on the page . My figure 1a also lacked connecting arrows between some of the boxes.

**We thank the Reviewer for their attention to detail, and have modified the figure to fix these issues.**

Suggested refs

Payne, M. R., A. J. Hobday, B. R. Mackenzie and D. Tommasi (2019). Editorial: Seasonal-to-decadal prediction of marine ecosystems: opportunities, approaches and applications. Frontiers in Marine Science: doi: 10.3389/fmars.2019.00100

Payne, M. R., A. J. Hobday, B. R. MacKenzie, D. Tommasi, D. P. Dempsey, S. M. M. Fässler, A. C. Haynie, R. Ji, G. Liu, P. D. Lynch, D. Matei, A. K. Miesner, K. E. Mills, K. O. Strand and E. Villarino (2017). Lessons from the first generation of marine ecological forecasts. Frontiers in Marine Science: doi: 10.3389/fmars.2017.00289

Reviewer: 2

Comments to the Author

This paper provides a quantitative review of the ecological forecasting literature, reflecting specifically on the degree to which “best practices” have been adopted and changes in those practices over time. In addition, the paper provides numerous insightful nuggets about the state of ecological forecasting, such as where we’re seeing the most forecasts (in space and by topic), the grain and extent of forecasts, the amount of data folks are using to calibrate data, and the trends in model skill with lead time for common forecast areas. Overall the paper is well written and accessible, the analyses are performed well, and the interpretation is sounds. I just have a few tiny suggestions, mostly focused on places where the authors could increase clarity.

**We thank the Reviewer for their positive feedback!**

Line 94: “forecastability” was used here, in the title, abstract, and few places later in the paper, but isn’t specifically defined anywhere. This seems like the best place to do so, and to provide a clear distinction between prediction, predictability, and forecastability (even without introducing “forecastability”, I see too many folk convolve the first two terms).

**We appreciate the Reviewer’s comments, and have provided an operational definition of forecastability as “realized forecast accuracy” in the Abstract (line 39) and Introduction (line 94–95). We also modified the text in this section to better illustrate the distinction between forecastability and predictability. This section now reads “Ecological forecasting can be a particularly powerful test of predictability, as forecasting requires predicting beyond the range of observed data (Dietze et al. 2018). Thus, comparisons of forecastability complement and extend existing theoretical and modeling-based work that has predicted how various factors (e.g., forecast horizon, computational irreducibility) influence the relative predictability of ecological variables (Beckage et al. 2011, Petchey et al. 2015)” (lines 99–103).**

Line 171-177: I think my biggest true critique of the paper is that the standardized matrix of questions that the authors used analyze these papers is never provided. Instead of putting this information in a supplement/appendix, as one normally would do, they instead put their scoring rubric inside a “forthcoming data publication” that is not accessible to reviewers. I came very close to sending the paper back at this point as non-replicable and noncompliant with ESA’s data policy. I also found it very odd that the matrix/rubric itself would be considered a data publication, while the dataset itself \*IS\* provided as a supplement. That said, it’s worth noting that that the datasets are likewise provided without any metadata (e.g. defining column names and units). All of these things should have been provided the reviewers and definitely need to be provided to the reviewers in the next revision.

**The “forthcoming data publication” in our initial submission was provided to Reviewers via an accessible link in the Open Data Statement as a provisional data product. This data publication contained all of the analysis data and metadata with column descriptions and units, as well as the R code used to analyze the data.**

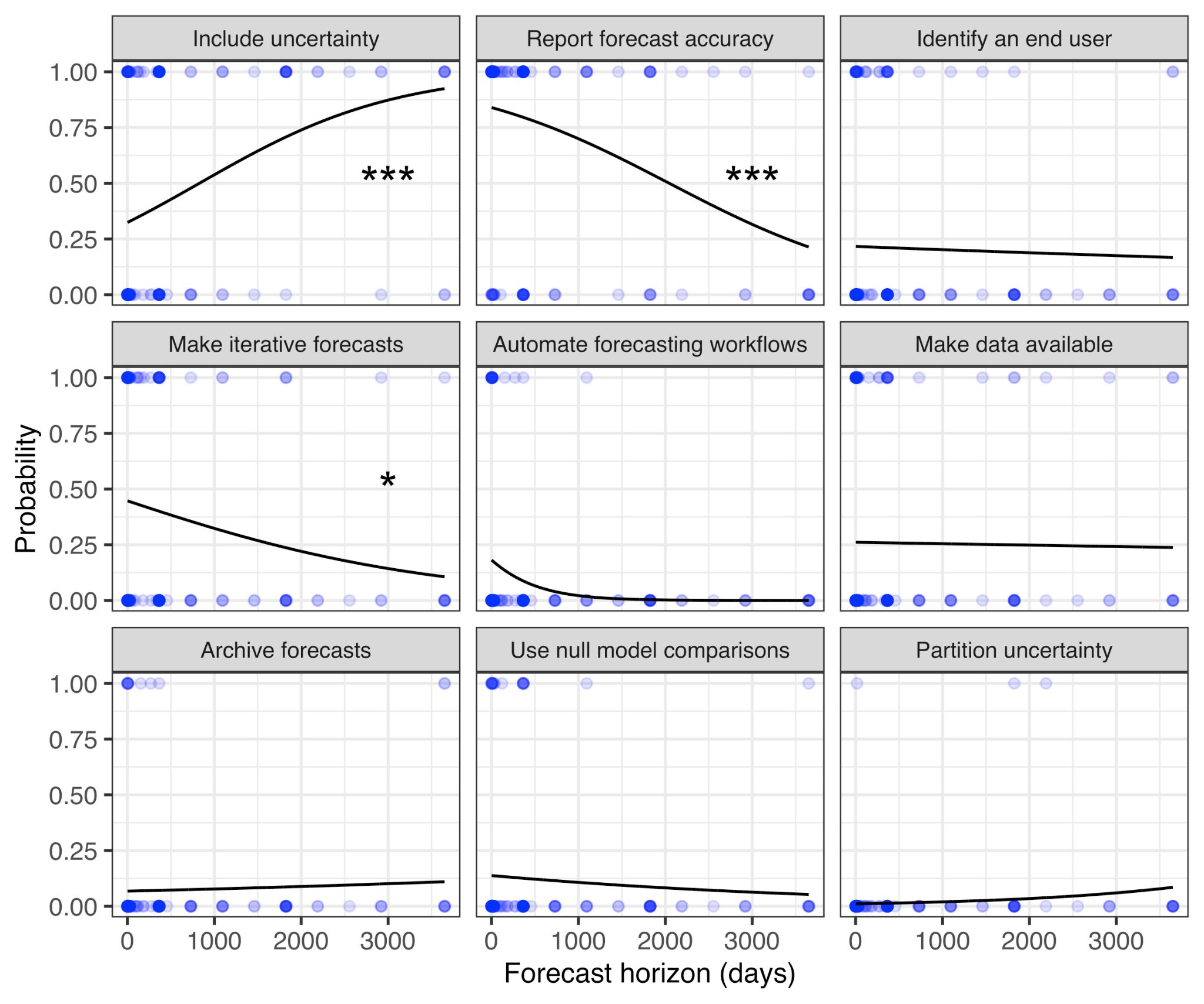
**As we did not receive any requested edits on this data publication from Reviewers, we have now published the dataset to the EDI repository so it has been assigned a permanent DOI (Lewis et al. 2021, reference provided above). We have changed all mentions of the “forthcoming data publication” to citations of this data publication. Additionally, to address the Reviewer’s concerns about the matrix/rubric itself, we have submitted an additional file for publication as supplemental information (Appendix S2). This file includes a list of the questions used to screen papers in the analysis to ease access to this information for readers.**

L180: Unclear what was actually done here. If reviewers screened papers individually, what were they checking with reviewers that didn’t screen that paper?

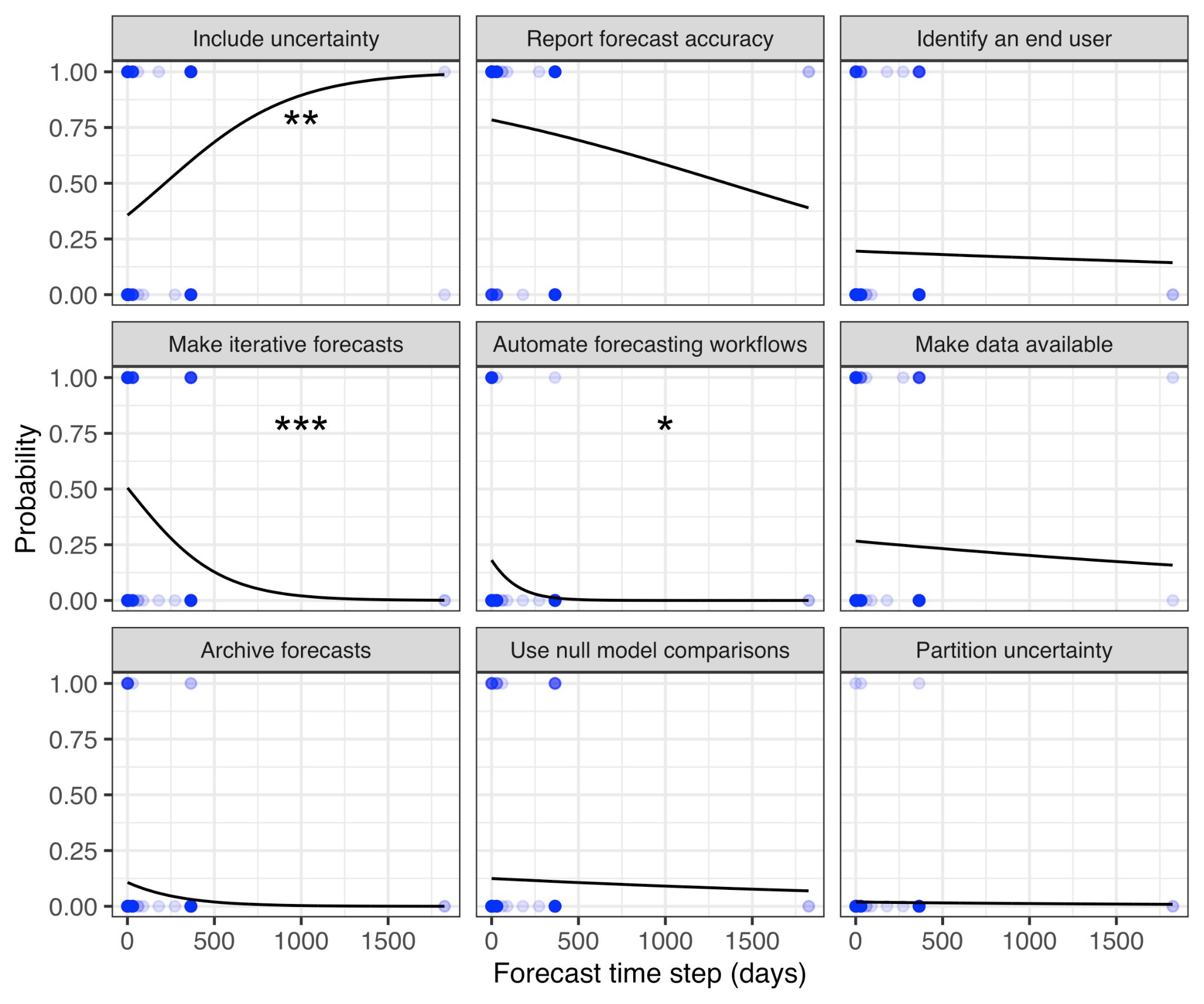
**Two reviewers first analyzed each paper independently (without consulting the other reviewer), and then second compared responses with each other. To clarify this point, we changed the word “individually” to “independently” on line 180.**

L216: Totally optional, but this section also made me curious how some of the best practices vary with forecast grain and limit (e.g. it wouldn’t surprise me if less frequent forecasts adhered to fewer of the best practices).

**We agree that this could be interesting! We did a preliminary analysis analyzing how best practice use changes with forecast horizon (Figure 1 below) and model time step (Figure 2). Intuitively, forecasts with longer horizons are less likely to be iterative and less likely to assess and report forecast accuracy, likely due to long latency before data become available to assess the forecast (Figure 1). Likewise, forecasts with long model time steps are less likely to be iterative or automated, for the same reason. Interestingly, the use of uncertainty increases with increasing forecast horizons and model time steps. While these figures are interesting and we appreciate the Reviewer’s suggestion, we have decided this analysis is beyond the scope of this manuscript. As more and more new forecasts become published, we are excited to further explore this topic in future work.**

****

**Figure 1: Best practice adoption relative to forecast horizon. Points demarcate whether or not an individual paper demonstrated the best practice (1 = observed, 0 = not observed), and lines represent logistic regression results. Significance of the year term in the regression is indicated using asterisks: \* indicates p < 0.05, \*\*\* indicates p < 0.001.**

****

**Figure 2: Best practice adoption relative to model time step. Points demarcate whether or not an individual paper demonstrated the best practice (1 = observed, 0 = not observed), and lines represent logistic regression results. Significance of the year term in the regression is indicated using asterisks: \* indicates p < 0.05, \*\* indicates p < 0.01, \*\*\* indicates p < 0.001.**

L263: The authors here state that “R2 is bias-corrected” but that’s not universally true. Some people calculate R2 as the square of r, some calculate it around a regression line between predictions and observations, and some calculate it as the variance explained by the forecast (i.e. around the predicted-observed 1:1 line instead of the bias-correction line). Thus it’s not universally true that all R2 are bias corrected (indeed, I’ve called out previous papers I’ve reviewed for presenting bias-corrected statistics when they should have been presenting the actual model skill). Was it true that 100% of the R2 values you came across were bias corrected? Did your matrix/rubric provide space for distinguishing these cases?

**We anticipate that all of the R2 values used in this dataset were bias-corrected. However, the method by which R2 was calculated was not always reported in papers, and was not recorded in our matrix analysis. To address the Reviewer’s comment, we clarified that R2 is typically (but not always) bias-corrected in the revised version: “While the fact that R2 is typically bias-corrected makes it an imperfect metric of forecast performance, it remains widely reported and uniquely suited to inter-study comparisons” (lines 265–267).**

L267: The rational for further restricting this analysis to a small majority of papers (56%) isn’t obvious. Please explain more about why this choice was made.

**We agree with the Reviewer that this was unclear in the original submission, and we have updated the text to clarify our motivation. The text now reads: “We selected all forecast variables that had at least three papers and three forecast horizons represented, and we plotted forecast performance (in R2) as a function of forecast horizon for these variables. To allow comparability between variables, we limited the analysis to forecast horizons between one and seven days, which were reported for all variables selected” (lines 269–273).**

L274: A rational for why this analysis was done using quantile regression, as opposed to standard linear regression, is not provided. Please explain the argument for this choice.

**We thank the Reviewer for their observation, and have added text to address this omission. The new text reads as follows: “Quantile regression was used rather than standard linear regression to account for heteroscedasticity and non-normal data distribution” (lines 279–281).**

L321: Here you say “all but one” but then you don’t tell us which one. You do eventually tell us this later, but that itself points to possible organizational issues and repetition with the Results.

**To address this comment, we added a parenthetical specification of the practice that has not increased in adoption over time. The sentence now reads: “All but one (“Use null model comparisons”) of our proposed best practices have been increasingly adopted over time” (lines 325–327).**

L345: While this stat is interesting, the crucial bit of missing information here is how many of the papers you reviewed were using DA in the first place. Does 67% represent 2 of 3 paper or 46 of 69 papers?

**We appreciate the Reviewer’s attention to detail, and have added the number of papers that used DA to this sentence: “67% of the 69 iterative forecasts only updated initial conditions” (line 357).**

Line 352-359: I find it odd that you present info about the trends in best practices (lines 349-352) BEFORE the more basic info presented here about their overall prevalence.

**We appreciate this organizational comment, and have restructured the three paragraphs that describe best practice adoption results to begin with statistics about overall prevalence before describing trends in adoption over time (lines 330–374).**

Line 358: of the papers that claim to be archiving forecasts on their own specific webpages, how many of those webpages actually still exist? How many contain accessible forecast archives (as opposed to just the current forecast)?

**This is an important point. We went back and checked and found that most were not available. We have added the following sentence to the results: “Only two of the seven papers that mentioned archiving forecasts on a website had links that were still functional as of 14 Jun 2021” (lines 369–370). We also added this this field of information to our data publication to ensure reproducibility.**

L359: If you’re down to n=3 papers, I think it would be useful to go ahead and cite them. Also, the separation between this text and line 351 about trends and line 355 about prevalence is an example of where the Results are bit disjointed/repetitive.

**We have added citations for the three papers that partition uncertainty (line 340), and we hope that our modification of the structure of this sentence following the Reviewer comment above has addressed this comment.**

L455-456: Great point!

**We appreciate the Reviewer’s enthusiasm and agree that research into the use and interpretation of forecast uncertainty has been important and productive!**

L602: EFI is a pretty big group, was this a specific working group or workshop?

**Multiple coauthors are involved with the Ecological Forecasting Initiative (EFI), and this paper has benefited from years of interactions and collaboration with EFI members at large, making it difficult to choose any one working group or workshop to reference. However, we do appreciate the Reviewer’s comment that EFI is a relatively large community, and we have modified the text to particularly highlight the contributions of the Theory working group and the RCN steering committee: “We thank members of the Ecological Forecasting Initiative (EFI), particularly the Theory working group and the Research Coordination Network (RCN) steering committee, for productive discussions throughout the process of writing this paper, and we thank two anonymous reviewers for their helpful revisions.” (lines 622–625).**