PEER REVIEW RECORD

The Squealer: Sensification of model exploration and model misfit

Andrew Gelman, Andrew H. Jaffe, Eliot Carlson and Philip Greengard

Vol. 1 (2026), Art. 16 · https://doi.org/10.68365/jrepstat.2026.16

The decision letters, the reports the authors read, their answers and the editor's signed decisions, as written. The reports come from the journal's review engine. The authors could have kept this record private until they confirmed the proof.

DECISION PATH

2026-10-26	Minor revision
2026-11-17	8 answers
2026-11-18	Accepted

Round 1: Minor revision

SIGNED DECISION

Confirmed and signed by Matthias Templ on 2026-10-26 · ORCID 0000-0002-8638-5276

DECISION LETTER

Thank you for your submission to JRepStat. We are inviting a minor revision. The reviewers' points follow, grouped by weight.

The paper introduces an interactive method for exploring model misfit by dragging a fitted curve toward user-specified pseudo-data points, re-fitting under a soft "gravitational" constraint, and reporting the resulting loss of fit through visual and auditory feedback. It will interest applied statisticians and domain modelers who want to interrogate why a fitted curve misses the data, demonstrated across golf putting, serial-dilution assays, CMB cosmology, and a Gaussian process. We would be glad to see the revised version.

How to read this letter. The decision rests on the items marked required. If a required item seems wrong to you, or cannot be done within this paper, decline it on the revision page and give the reason. A reasoned decline is a legitimate answer, and the editor weighs it. Reviewers can be mistaken. Items marked recommended or optional are advice from the reviewers. Adopt them, decline them with a sentence, or leave them for later work. None of that counts against the paper. The reviewer reports below carry the same marks. The editor who signs this decision may weigh an item differently in the accompanying message, and that message prevails.

REQUIRED (the decision rests on these):

- [writing-quality] ` Several word-level errors that are not extraction artefacts should be corrected: "1.25 million attempted puts" should be "putts"; "the final bin cantered at 75 feet" should be "centered"; "forcing the curve go through" is missing "to"; "adjusting the posterior the posterior so that" repeats "the posterior"; and "The data binned by distance from the hole, binned into 31 categories" is missing a verb and repeats "binned". These change or obscure meaning, so the authors should decide the correct wording.
- [methods-soundness] In Section 1.3 the augmented posterior is written p(θ|*y*) ∝ *p(θ|y) normal(y|*x, *σ*); the normal density should be conditioned on the curve, normal(y|*g(x|*θ), σ*), as given correctly in Section 1.2. Correct the density argument so the augmented target is unambiguous.
- [Reviewer 1] Formula (1) defines the posterior sd of g(x*) as the posterior average of the conditional sds, which is not the sd of the posterior of g. Correct the integral or state it as an approximation.

- [Reviewer 1] The golf interpretation rests on the claim that short-distance bins contain failed attempts from longer putts. Justify this from the data structure or qualify it.
- [Reviewer 2] Equation (1) defines σ* incoherently: the integrand sd(g(x|*θ)) is zero for fixed θ because g is deterministic given θ, the differential reads dy where θ is the variable of integration, and a posterior sd cannot equal a posterior average of per-θ sds. State the intended quantity (the posterior standard deviation of g(x)*).
- [Reviewer 2] The quantity conveyed by sound is described two ways. Figure 1 and Figure 2 captions use the change in log posterior, log p(θ*|x,y) − log p($\hat{\theta}$|x,y), while the Section 2.1 prototype text uses the decrease in log likelihood under the original model. Reconcile which quantity drives the squeal and use it consistently.
- [Reviewer 2] Equation (1) is offered as the default σ* rule but is not the rule used in the GP example (fixed 10^{-6}) or the CMB example ($\bar{\sigma}_\ell$*), and the squeal magnitude (e.g. cost 150.4 vs 20.2) has no absolute scale because it depends on σ*. Either add a short treatment of what the squeal magnitude means and why σ* differs by example, or qualify eq (1) as one default and the squeal as a relative diagnostic.
- [Reviewer 3] Post runnable code and data for all four examples, not only the golf prototype, each with a command that regenerates its figure. The paper's four demonstrations are its evidence, and only one has a linked artifact.

RECOMMENDED:

- [software-landscape] The interactive prototype at the GitHub link is a manipulable-graphics system in the same space as GGobi (already cited), loon, plotly and Shiny. Please add one or two sentences placing the dragging-and-updating interface relative to existing interactive graphics systems, so a reader can see what the Squealer adds beyond brushing and linked views.
- [software-landscape] The Gaussian process example and the dilution model are fit with the authors' own Stan and closed-form R code. Since these serve only to illustrate the method, no cross-check is required, but one line naming an established GP implementation (kernlab, GPfit, or GPy/GPflow) alongside the Rasmussen and Williams citation would help readers who want to reproduce the GP fit with a standard tool.
- [statistical-rigor] The predicted "approximately 80% shift toward the pseudo-data point" is derived from a normal, locally linear baseline (the ratio $1^2/(1^2+0.5^2)$), then stated as a general expectation. The example models (the golf arcsine-normal curve, the four-parameter logistic assay, the CMB spectrum, the GP) are nonlinear, so the actual shift will depart from 80%. Add one sentence stating that this figure is a normal-linear approximation and that realized shifts in the nonlinear examples differ.
- [statistical-rigor] The "cost" of each perturbation (for example 150.4, 40.0, 20.2 in Figure 10, and the hyperparameter changes in Table 1) is reported as a single log-likelihood decrement with no indication of sampling or Monte Carlo uncertainty. For a tool whose purpose is to quantify discomfort with a fit, a reader cannot tell which differences are meaningful. State whether these costs are deterministic given the fitted mode or carry Monte Carlo error, and if the latter give a rough magnitude.
- [statistical-rigor] The ±1 standard error intervals shown for log β2 and log σy in Figure 9 are, for the two perturbed samples, shrunk by the pseudo-data mechanism rather than by the data, as the text notes in Section 1.4 and 2.2. Readers may still read those narrowed intervals as genuine posterior precision. Mark in the Figure 9 caption that the intervals for samples 3 and 23 are artifacts of the gravitational constraint and do not represent real uncertainty reduction.
- [visualisation] In Figure 10 the best-fit curve is red and the Squealer curve is orange, with blue data error bars. Red against orange is hard to separate for readers with protan or deutan colour vision, and these are the two curves the reader must compare. Please separate them more, for example a solid dark line for the best fit and a distinct dashed line plus a non-orange hue for the Squealer curve, so the contrast survives in greyscale. The red/blue coding in Figures 1, 2, 5, 6 and 8 reads more safely, since dark blue, light blue and red stay distinguishable.
- [visualisation] Figure 9 places the two panels on different axis scales and notes this in the caption. The side-by-side layout invites a direct visual comparison that the differing scales defeat, since the apparent spread of points is not comparable across panels. Consider a shared x-axis range, or an explicit marker of the common region, so the shift in samples 3 and 23 reads against a fixed frame.

- [visualisation] The central dashboard panels in Figures 1, 2, 5 and 6 carry many small stacked labels (theta[1] through theta[6], log prior, log lik, log post) and the right-hand residual panels use axis ticks like "−1e-05 1e-05 3e-05" in small type. At print size these are hard to read. Larger tick labels and a short key for the dashboard rows would help, given the paper leans on these panels to carry its argument.
- [methods-soundness] For point summaries other than the mode the paper states there is no guarantee the log posterior decreases, so the quantity to be sonified is undefined, and the KL alternative is proposed but not applied in any example. Please state in one sentence which scalar is actually used for the auditory signal in each example (the examples appear to use the mode or the fixed-hyperparameter closed form).
- [Reviewer 1] The abstract describes sound that grows "louder and more unpleasant" while the prototype maps discrepancy to pitch. Reconcile these and add a sentence on the rationale for the sonic mapping.
- [Reviewer 1] The Figure 11 caption states hyperparameters are fixed, while Table 1 shows them changing across conditions. Clarify what is held fixed and when.
- [Reviewer 2] Report how far the curve moves and how large the squeal is as σ^* varies, for at least the golf and GP examples, so a reader can judge the sensitivity of the diagnostic to this tuning choice.
- [Reviewer 3] Capture the compute environment (Stan, R or Python, optimizer library versions) and the seeds for the simulated GP example and the posterior draws, with a short determinism note.
- [Reviewer 3] Document the sound mapping used by the prototype (how pitch and loudness follow the change in log likelihood) so the sonification is reproducible, not only the curve-fitting.

OPTIONAL:

- [Reviewer 1] A small user study of whether the auditory channel aids misfit understanding would support the claims about non-expert intuition.

We will correct these in production, and you confirm them in the proof:

- [consistency-check] Section 2.2 states Figure 7 shows "22 unknown samples", but the rest of the paper uses 23: samples are numbered 1 to 23 in Figure 9, the text refers to "samples 23 and 3", "the set of 23 values of the parameter β2", and Figure 9′s caption says "23 unknown samples". The "22" contradicts the count the paper otherwise uses and should be reconciled to the correct value.
- [consistency-check] The Figure 8 caption reads "Applying the Squealer to the data and model from Figure 8", referring to itself. From the text (Section 2.2 introduces the plate and curves in Figure 7) the intended reference is Figure 7.
- [consistency-check] Section 1.1 refers to "Figure 2.2" when discussing the four-parameter logistic assay model ("as discussed further in Figure 2.2"). No Figure 2.2 exists; the intended target is Section 2.2.
- [consistency-check] In Section 2.1 the full paragraph describing the software prototype ("We include a software prototype with the golf example at https://github.com/elc45/squealer-prototype … responding in real time.") is printed twice in succession. One copy should be removed.
- [methods-soundness] In Eq (1) the averaging integral is written with measure dy but integrates $p(\theta|y)$ over θ, so the measure should be $d\theta$. (the correct form is fixed by the surrounding text)
- [methods-soundness] The software-prototype paragraph in Section 2.1 is duplicated verbatim.
- [prose-and-style] Section 2.4, "adjusting the posterior the posterior so that it moves through the pseudo-data point" repeats "the posterior".

REVIEWS

Report Reviewer 1, recommends minor

Written by the journal's review engine in the role of the application reviewer, and read by the editor before the decision. How the engine works is on the peer review page.

This is an appealing and well-motivated contribution. The central idea, inverting the usual direction of exploration so that the user moves the curve and lets the parameters follow under a soft constraint, matches how applied modelers actually think about misfit. The framing in terms of pulling a fitted curve toward pseudo-data, then reading off the statistical price in the likelihood and in individual data points, is

a genuine addition to the exploratory toolkit. The four examples are chosen well and span a useful range, from a tightly constrained two-parameter model to a nonparametric GP, and each one teaches something about why a model resists the fit the eye wants.

My largest application concern is the golf narrative, because it carries the paper's cleanest interpretive payoff. The explanation for the large likelihood cost is that the short-distance bins hold "a pileup of data near x = 0" made up of failed attempts from other distances that came close. In the aggregate putting data I am aware of, each attempt is recorded at its true starting distance, so a missed long putt does not re-enter a short bin. If the data here are structured differently, say by recording a follow-up putt, that needs stating, because the whole point of the panel depends on where those hundreds of thousands of attempts actually come from. Please either justify this from the data definition or qualify the sentence so the conclusion rests on bin counts rather than on a claimed re-binning of misses.

On the method itself, equation (1) sets the gravitational scale to half the posterior sd of g(x*), then writes that quantity as the posterior average of the conditional sds of g(x*|theta). Those are not equal in general, since the sd of a mixture is not the mixture of the component sds, and the displayed integral is also written with dy where dθ is meant. For a method paper this formula is load-bearing, so it should be corrected or presented explicitly as an approximation with a note on when it is adequate.

On sonification, which is where the title places the paper, the sonic layer is described rather than demonstrated, and the manuscript says as much. That is a defensible scope for a method paper. Two things would still help a reader from the sonification side. The abstract promises a sound that grows louder and more unpleasant, while the golf prototype maps the discrepancy to pitch, and these are different perceptual channels with different detectability properties. A sentence reconciling the two, and a sentence on why the chosen mapping was expected to work, would ground the design in the auditory-display literature the introduction already gestures at. The claim that the tool builds intuition for non-experts is reasonable as a motivation and is currently unsupported by any use evidence, which is fine to leave as future work if stated that way.

The bioassay and cosmology examples read well, and the honest treatment of the posterior-constriction artifact in Section 1.4, carried through to the shrunken sample 3 and 23 uncertainties in Figure 9, is the right way to handle a known limitation. In the GP example, the caption says the kernel hyperparameters are fixed while Table 1 reports them moving across the zero, one, and two pseudo-point fits. I take it the hyperparameters are re-fit and then held at their new values for sampling, but the reader should not have to reconstruct that.

A few cross-references and a repeated paragraph need the production pass noted above. None of these affect the result.

Numbered suggestions to the authors:

1. REQUIRED Correct equation (1): the posterior sd of g(x*) is not the posterior average of the conditional sds sd(g(x*|theta)), and the integral is written over dy rather than dθ. State the exact quantity, or present (1) as an approximation with a note on when it holds.
2. REQUIRED Justify the claim in Section 2.1 that the short-distance bins include failed attempts from longer putts, from the definition of the putting data, or qualify the interpretation so it rests on the large bin counts near x = 0 rather than on a re-binning of misses.
3. [editor_fix] Remove the duplicated software-prototype paragraph in Section 2.1.
4. [editor_fix] Fix the Figure 8 caption, which refers to "Figure 8" where Figure 7 is meant, and the text reference "Figure 2.2", which should be Section 2.2.
5. RECOMMENDED Reconcile the sound description in the abstract (louder and more unpleasant) with the pitch mapping in the prototype, and add one or two sentences on the rationale for the chosen sonic mapping with reference to the auditory-display work cited in Section 1.5.
6. RECOMMENDED Clarify the Figure 11 caption against Table 1: state which hyperparameters are re-fit and which are held fixed for sampling at each stage.
7. RECOMMENDED Confirm in the paper that data and code for all four examples are posted, not the golf prototype alone.

8. OPTIONAL A small study of whether listeners detect the intended misfit from the audio channel would support the claim that the tool aids non-expert intuition.

Report Reviewer 2, recommends minor

Written by the journal's review engine in the role of the methodology reviewer, and read by the editor before the decision. How the engine works is on the peer review page.

This is a clearly written methods-and-ideas paper. The central device is appealing and, as far as I can tell, sound in its basic logic: multiply the posterior by a normal "gravity" term centred on a user-placed pseudo-data point, re-optimise, and read off how much the fit resists the pull. Treating the squeal as the model's resistance to a desired curve shape is a reasonable exploratory diagnostic, and the four examples cover a useful range of structures, from a two-parameter curve to a nonparametric GP. The precision-weighting argument for the roughly 80% shift ($1/(1 + 0.5^2) = 0.8$) is correct, and the honesty about the method's limits in Sections 1.3 and 1.4 is welcome. My comments are mostly about tightening the statistical statements, because a method paper is judged on whether its formulas and its reported diagnostic mean what they say.

The definition of the gravity scale in equation (1) does not hold as written. The right-hand side sets the posterior standard deviation of g(x*) equal to one half of a posterior average of sd(g(x|θ))*. For fixed θ the curve value g(x|*θ) = E(y|x,*θ) is a single number, so its standard deviation is zero, and the integral is written against dy rather than the posterior over θ. The intended object is the posterior standard deviation of g(x*) across draws of θ, which is what the surrounding text and Figure 3 clearly use. Please restate eq (1) so the symbol matches that quantity.

A second clarity point concerns the quantity that squeals. The captions of Figures 1 and 2 map sound to the change in log posterior density, while the prototype description in Section 2.1 maps it to the decrease in log likelihood under the original model. These coincide only when the prior contribution does not change, which is exactly the hierarchical case (dilution assays) where it does change, since the paper itself states the penalty there is paid in the prior. Pick one quantity and carry it through the captions, the prototype text, and the discussion.

The deeper methodological question is how to read the squeal. Equation (1) is presented as the default scale, but the GP example pins the pseudo-point with variance 10^{-6} and the CMB example uses its own $\bar{\sigma}_\ell$, *so the default rule is used in at most two of the four cases. Because the reported cost scales with* σ, the numbers in Figure 10 (150.4, 40.0, 20.2) are interpretable only relative to one another within a fixed σ*, and the same is true of the golf change-in-log-likelihood plot. The paper would be on firmer ground if it said this plainly: the squeal is a relative diagnostic, eq (1) is one default, and different problems warrant different scales. A sensitivity curve for the golf and GP cases would make the point concrete without new data.

On reproducibility, the prototype covers the golf case; the dilution, CMB, and GP results, including the hyperparameters in Table 1, are the ones I would most want to rerun, since they involve real fitting choices. Posting the code for all four would let a reader verify the figures. Two production-level items: the prototype paragraph in Section 2.1 is duplicated verbatim, and the Figure 8 caption points at Figure 8 where it means Figure 7.

Numbered suggestions to the authors:

1. REQUIRED Correct equation (1) so σ* is defined as (one half of) the posterior standard deviation of g(x*) over draws of θ. As written the integrand is zero for fixed θ, the differential reads dy, and a posterior sd is set equal to a posterior mean of per-θ sds.
2. REQUIRED Use a single quantity for the squeal throughout. The Figure 1 and 2 captions use the change in log posterior; the Section 2.1 prototype text uses the decrease in log likelihood. These differ exactly in the hierarchical dilution case where the prior term moves.
3. REQUIRED Say what the squeal magnitude means and why σ* is chosen differently across examples (eq (1) default, 10^{-6} in the GP, $\bar{\sigma}_\ell$* in the CMB), or add one or two sentences qualifying eq (1) as a default and the squeal as a scale-dependent relative diagnostic.
4. [editor_fix] Remove the duplicated prototype paragraph in Section 2.1.
5. [editor_fix] Change the Figure 8 caption reference from Figure 8 to Figure 7.

6. RECOMMENDED Post runnable code for all four examples so the dilution, CMB, GP figures and Table 1 can be reproduced, not only the golf prototype.
7. RECOMMENDED Show how the curve shift and the squeal change as σ^* varies for the golf and GP examples, so the sensitivity of the diagnostic to this tuning choice is visible.

Report Reviewer 3, recommends major

Written by the journal's review engine in the role of the software reviewer, and read by the editor before the decision. How the engine works is on the peer review page.

I read this as a software and reproducibility reviewer. My questions are whether I can run the thing, whether the code is sound, and whether the figures in the paper are the code's output. The method itself is clearly described and the idea is attractive. The paper is also plain that it demonstrates principles with static graphics and that the dynamic, sonic, and haptic forms are future work, so I am not holding it to a standard of a shipped interactive tool.

The central gap for me is coverage of the artifacts. The abstract and body present four worked examples (golf, dilution assays, CMB power spectrum, GP temperature), and each figure is an output of code. Only the golf example has a linked prototype, at the GitHub URL in Section 2.1. The footnote says all data and code "will be freely available and posted on Github", in the future tense. For a new-method paper whose argument rests entirely on these four demonstrations, the reproducibility minimum is code plus a command that regenerates each figure, available at submission. Please post the code and data for all four examples, each with a runnable entry point. I could not inspect the repository contents, so I am asking only that the artifacts exist and run, not for anything about their internal structure.

Two formula statements are wrong as printed and should be fixed because a reader reconstructing the method from them would be misled. In Section 1.3 the augmented posterior is written $p(\theta|y) \propto p(\theta|y)\, normal(y|x, \sigma)$; the normal density should be centered at $g(x^*|\theta)$, matching the construction in Section 1.2. In equation (1) the integral averages over the posterior but is written with dy where it should be $d\theta$. Both are quick corrections.

A few things fall to the production list rather than to you as analysis. Section 2.1 contains a full paragraph printed twice, the block beginning "We include a software prototype with the golf example" and ending at the L-BFGS sentence. The Figure 8 caption refers to "Figure 8" where it means Figure 7, and the Figure 1 discussion writes "Figure 2.2" for what is Section 2.2.

On reproducibility beyond the minimum, two additions would let the journal verify the examples and would document what you already have. The GP temperature case uses simulated data, fitted hyperparameters, and posterior draws, and the fitted values in Table 1 depend on the random draws and the optimizer. Recording the environment versions (Stan, R or Python, the optimizer library) and the seeds, with a one-line determinism note, would make Table 1 and Figure 11 reproducible. Separately, the sonification is the paper's headline move, and right now only the curve fitting is described well enough to repeat. A short statement of how pitch and loudness map to the change in log likelihood in the prototype would make the sound reproducible too.

Numbered suggestions to the authors:

1. REQUIRED Post runnable code and data for all four examples, not only the golf prototype, each with a command that regenerates its figure.
2. REQUIRED Correct the augmented-posterior density in Section 1.3 to be centered at $g(x^*|\theta)$, and change the integration variable in equation (1) from dy to $d\theta$.
3. [editor_fix] Remove the duplicated paragraph in Section 2.1 (the "We include a software prototype … L-BFGS optimizer" block printed twice).
4. [editor_fix] Fix the Figure 8 caption to reference Figure 7, and change "Figure 2.2" in the Figure 1 discussion to Section 2.2.
5. RECOMMENDED Capture the compute environment versions and the seeds for the simulated GP example and the posterior draws, with a short determinism note, so Table 1 and Figure 11 can be reproduced.
6. RECOMMENDED Document the prototype's sound mapping (how pitch and loudness follow the change in log likelihood) so the sonification can be reproduced alongside the curve fitting.

REVISION, RECEIVED 2026-11-17

Authors confirm each required point as done or give their reason for declining it. Further answers are voluntary, because the review compares every revision with the previous version and the next round's reports say what remains open.

> REQUIRED ` Several word-level errors that are not extraction artefacts should be corrected: "1.25 million attempted puts" should be "putts"; "the final bin cantered at 75 feet" should be "centered"; "forcing the curve go through" is missing "to"; "adjusting the posterior the posterior so that" repeats "the posterior"; and "The data binned by distance from the hole, binned into 31 categories" is missing a verb and repeats "binned". These change or obscure meaning, so the authors should decide the correct wording.

Done.

> REQUIRED In Section 1.3 the augmented posterior is written $p(\theta^*|y) \propto p(\theta|y)$ normal$(y^*|x^*, \sigma^*)$; the normal density should be conditioned on the curve, normal$(y^*|g(x^*|\theta), \sigma^*)$, as given correctly in Section 1.2. Correct the density argument so the augmented target is unambiguous.

Done.

> REQUIRED Formula (1) defines the posterior sd of g(x*) as the posterior average of the conditional sds, which is not the sd of the posterior of g. Correct the integral or state it as an approximation.

Done.

> REQUIRED The golf interpretation rests on the claim that short-distance bins contain failed attempts from longer putts. Justify this from the data structure or qualify it.

Done.

> REQUIRED Equation (1) defines σ* incoherently: the integrand sd(g(x*|θ)) is zero for fixed θ because g is deterministic given θ, the differential reads dy where θ is the variable of integration, and a posterior sd cannot equal a posterior average of per-θ sds. State the intended quantity (the posterior standard deviation of g(x*)).

Done.

> REQUIRED The quantity conveyed by sound is described two ways. Figure 1 and Figure 2 captions use the change in log posterior, $\log p(\theta^*|x,y) - \log p(\hat{\theta}|x,y)$, while the Section 2.1 prototype text uses the decrease in log likelihood under the original model. Reconcile which quantity drives the squeal and use it consistently.

Done.

> REQUIRED Equation (1) is offered as the default σ* rule but is not the rule used in the GP example (fixed 10^{-6}) or the CMB example ($\bar{\sigma}_\ell^*$), and the squeal magnitude (e.g. cost 150.4 vs 20.2) has no absolute scale because it depends on σ*. Either add a short treatment of what the squeal magnitude means and why σ* differs by example, or qualify eq (1) as one default and the squeal as a relative diagnostic.

Done.

> REQUIRED Post runnable code and data for all four examples, not only the golf prototype, each with a command that regenerates its figure. The paper's four demonstrations are its evidence, and only one has a linked artifact.

Done.

Round 2: Accepted

SIGNED DECISION

Signed by Matthias Templ on 2026-11-18 · ORCID 0000-0002-8638-5276

Accepted. All three reviewers recommend acceptance or a minor revision of the revised paper. The authors supplied code and data for all four examples, and the journal ran them: every figure regenerates, with small differences in Table 1 and the CMB costs. Open points go to the authors as suggestions with the proof.

DECISION LETTER

Congratulations, your manuscript has been accepted for publication in JRepStat. Production begins now.

Where it is not on file yet, we still need the following for production.

- a valid ORCID iD for every author

- a captured computational environment (dependencies/lockfile/container)
- a signed determinism attestation for the entry-point command

REVIEWS

Report Reviewer 1, recommends accept

Written by the journal's review engine in the role of the application reviewer, and read by the editor before the decision. How the engine works is on the peer review page.

Of the four must-fix points and three lighter ones I raised, all of the required and editor-fix items are met and nothing material remains open.

Equation (1) now reads as the posterior variance of g(x*), the integral is over dθ, and the formula is no longer the average of conditional standard deviations. This was the load-bearing correction and it is done. On the golf example, the panel's conclusion now turns on the large bin counts near x = 0 ("hundreds of thousands of putt attempts … hence the binomial likelihood is very strong there"), with the pileup of short putts offered as the reason those bins are large. That is the right ordering, and for ShotLink-style data the pileup of short second putts after missed long putts is a defensible account of the counts. The phrasing about failed attempts "from other distances" is still slightly loose, since the follow-up putt rather than the original miss is what lands in the short bin, but the statistical point the panel depends on holds, so I would not hold the paper for it.

One recommended item remains the author's choice. The repository footnote is still written in the future tense, so a reader cannot yet tell whether data and code for all four examples are posted or only the golf prototype. A one-word change to the present tense, once the artifacts are up, settles it. This does not bear on the recommendation.

The remaining recommended items from the prior round (reconciling the abstract's "more unpleasant" sound with the prototype's pitch mapping, and clarifying the Figure 11 caption against Table 1) would still strengthen the paper, and they are optional strengthenings rather than anything the result rests on.

Numbered suggestions to the authors:

1. RECOMMENDED Change the repository footnote to the present tense once the data and code for all four examples are posted, so a reader can confirm the full set is available rather than the golf prototype alone.

Report Reviewer 2, recommends accept

Written by the journal's review engine in the role of the methodology reviewer, and read by the editor before the decision. How the engine works is on the peer review page.

Of the three required points and two editor fixes I raised, equation (1) and both caption/reference fixes are met. Equation (1) now reads as the posterior standard deviation of g(x*), integrated against p(θ|y) dθ with* $\bar{g}(x)$ the posterior mean, so the earlier problems (the dy differential, a per-θ sd that is zero, a posterior sd equated to a mean of per-θ sds) are gone. The duplicated prototype paragraph is removed and the Figure 8 caption now points at Figure 7.

Two residuals remain, and neither blocks the paper. The squeal quantity is still described as the change in log posterior in the Figure 1 caption (p. 2) and as the decrease in log likelihood in the Section 2.1 prototype text (p. 7). For the golf prototype the two nearly coincide, since that two-parameter fit with a million observations barely moves the prior term, so the choice is defensible. The paper does not say so, and a reader meeting both descriptions will wonder which drives the sound. One sentence noting the prototype uses the likelihood change because the prior contribution is negligible there would close it.

On the scale question, equation (1) is now framed as "a reasonable first choice," which does present it as one default among possible σ* rules, and that addresses the main part of my earlier point. What is still absent is a single sentence saying the cost numbers are comparable only within a fixed σ*. The CMB costs in Figure 10 (150.4, 40.0, 20.2) read naturally as relative to one another, and stating that explicitly would keep a reader from treating them as absolute.

Numbered suggestions to the authors:

1. RECOMMENDED In Section 2.1 (p. 7), add one clause reconciling the prototype's log-likelihood squeal with the log-posterior squeal in the Figure 1 caption (p. 2), for example that the prior term is negligible in the golf fit so the two agree there.
2. RECOMMENDED Add one sentence near Section 1.2 (p. 3) or the Figure 10 discussion (p. 14) stating that the squeal magnitude is comparable only within a fixed σ^*, so the reported costs are a relative diagnostic.

Report Reviewer 3, recommends minor

Written by the journal's review engine in the role of the software reviewer, and read by the editor before the decision. How the engine works is on the peer review page.

Of the two required points and two production fixes I raised, the three I can verify in the text are met: equation (1) now integrates over θ, the Section 1.3 augmented posterior is centered at $g(x|\theta_*)$, the duplicated prototype paragraph in Section 2.1 is gone, and the Figure 8 caption and the "Section 2.2" cross-reference are correct.

One item is still open, and it is the one I flagged as the central gap. The availability footnote on page 1 remains in the future tense ("will be freely available and posted on Github"), and the golf prototype URL in Section 2.1 is still the only linked artifact. The response marks the required coverage point as done, and I cannot see the repository to confirm what is now in it. If the code and data for the dilution assay, CMB, and GP examples are posted, the fix is to update the footnote and add the links so a reader can find them. Until the manuscript points to them, the paper's four demonstrations still read as having one reproducible artifact.

The one recommended point I can check is not addressed: the GP temperature example (page 14) still gives no seeds or library versions, so Table 1 and Figure 11 cannot be reproduced exactly. This was voluntary and does not hold the paper. The sound-mapping description is now adequate for the prototype, since Section 2.1 states that pitch is proportional to the decrease in log likelihood.

Numbered suggestions to the authors:

1. REQUIRED Update the availability footnote on page 1, which still reads "will be freely available and posted on Github" and links only the golf prototype, to link the posted code and data for all four examples, each with a command that regenerates its figure.
2. RECOMMENDED Record the seeds and environment versions (python, torch, optimizer) for the GP example on page 14, with a one-line determinism note, so Table 1 and Figure 11 can be reproduced.