

Response Letter to Reviewers of the Manuscript “Global Long-wavelength Body-wave Catalog (GLoBoCat) for Lower Mantle P and S Wave Travel Times”

Jun Su*, Christine Houser, John W. Hernlund

*junsu@jamstec.co.jp

September 18, 2026

We sincerely thank the reviewers for their careful reading of our manuscript and for their constructive criticisms and insightful suggestions. In response, we have performed additional work to check the data and improve quality control, revised the manuscript to provide more cautious and precise interpretation of the results, to clarify the presentation of key arguments, and to improve the overall readability and organization of the text. We have addressed each comment in detail, as outlined below, and have implemented corresponding changes in the revised version. We hope that these revisions adequately address the reviewers’ concerns and further enhance the quality of the manuscript for the benefit of its readers.

Reviewer 1

Su et al. present a machine-learning approach for measuring arrival times of first-motion P waves (Z component) and S waves (T component). The main outcome is GloBoCat, a catalogue of traveltimes picks that is roughly an order of magnitude larger than the global long-period traveltimes catalogue produced by Houser and colleagues nearly two decades ago. Consistent with prior work, the authors observe that the ratio $\delta Ts/\delta Tp$ of direct P- and S-wave traveltimes anomalies increases with epicentral distance beyond 80° . They infer that the ratio $R = \delta \ln Vs/\delta \ln Vp$ of mantle velocity anomalies increases with depth in the lowermost mantle, under the assumption that structure near the turning points dominates the observed traveltimes anomalies. The direct comparison between ISC and GloBoCat traveltimes anomalies reveals systematic differences but scatter is large. The hypothesis that these differences reflect physical dispersion (via wave-front healing) is interesting, but it is not yet explored in sufficient depth to support the interpretation.

My main comments are below. I have also uploaded the manuscript with annotations using Adobe Acrobat. I hope these suggestions will help strengthen the manuscript.

1) Motivation and positioning of the machine-learning approach.

The Introduction could motivate the machine-learning approach more strongly and more specifically in relation to prior “automatic” or semi-automatic approaches. In my experience, a fully automatic picker based on waveform correlation can be fast and robust, provided the signal-to-noise ratio is higher than 1 and a set of quality-control rules is followed (tuned for teleseismic long-period body waves, $T > 15$ s; see Ritsema et al., 2002). It seems plausible that the approaches of Bolton and Masters (2001) and Houser et al. (2008) could be adapted to run in a more fully automatic mode.

I have found that the bottlenecks were not the picking itself, but (i) downloading data at scale, (ii) fixing metadata, and (iii) reconciling differences among source parameters (e.g., CMT vs USGS/NEIC) and their effects on waveforms. Picking P and S waves at regional distances and/or in band-limited data can be difficult, but I do not understand the statement that “rule-based automatic picking ... underperform human experts picking real global long period data” (lines 56-58). What are the issues? For example: phase misidentification, sensitivity to source complexity, emergent

onsets, polarity changes, interfering phases, instrumentation issues, or inadequate QC rules? Making this explicit would clarify how the Transformer helps beyond automation.

We agree that the usage of “underperform” was too general, and we have rewritten this portion of the Introduction. The revised text states explicitly the two obstacles that rule-based automated pickers face with global long period data: (1) their picking functions are sensitive to variable noise sources, so that thresholds and filter parameters must be re-tuned for individual stations and events (Zhang et al., 2018, GJI, now cited); and (2) no single set of rules generalizes across the highly heterogeneous global networks and source conditions—for example, onset detectors relying on first-motion polarity fail when a reflector near the source reverses the apparent polarity. Human experts overcome these obstacles by selecting and adjusting a measurement strategy only after synthesizing the source mechanism, path complexity, and record quality, a judgment that is difficult to encode as fixed rules. This is precisely how the Transformer helps beyond automation: it learns this judgment from the training picks rather than from prescribed rules.

We have also added our motivation relative to prior semi-automatic approaches to the Introduction (L57–69). We fully agree that downloading data at scale, fixing metadata, and reconciling source parameters are equally serious bottlenecks; our pipeline addresses these with the same processing procedures as Houser et al. (2008), and the revised Introduction now motivates the choice of machine learning specifically for the picking stage.

2) Clarify “Earthquake Transformer” for a geophysical audience.

For readers unfamiliar with the Earthquake Transformer framework, the opening paragraph of Section 2 is hard to parse. Terms such as “weights and connections”, “different aspects and nuances of the input signal”, “global attention layer”, “intricate patterns and relationships”, and “binary cross-entropy” may be standard in ML, but I suspect they are not self-explanatory to many seismologists. I suggest rewriting this paragraph with a minimal, seismology-facing description. A short schematic might help, but even a concise plain-language explanation would improve readability.

We have rewritten the opening of Section 2 for a seismological audience. The revised text replaces the ML jargon in the Earthquake Transformer summary with seismology-facing descriptions (L96–125). The architectural details required for reproducibility (number of blocks, parameter counts, sampling rate, and probability width) are retained.

3) Definition and physical meaning of the “onset time”.

Line 100 refers to “onset times hand-picked by manual cross-correlation with synthetic waveforms.” This is confusing. “Onset” usually refers to the first departure from the noise level, whereas cross-correlation is performed over a waveform segment and yields a time shift relative to a reference, not necessarily the earliest detectable onset. Maybe a schematic could help clarify this? It matters for the paper’s later discussion of dispersion and for whether the measurement can be interpreted with ray-based sensitivity concentrated near turning points. If the measurement is effectively a band-limited time shift of a finite-length waveform, the sensitivity may be more distributed than a ray.

We agree that placing the terms “onset” and “cross-correlation” together in the original sentence made the measurement procedure appear contradictory. The sentence was intended as shorthand for the two-step procedure developed by Houser et al. (2008) and continued in the unpublished work used here, which expanded the dataset of Masters et al. (2000) by introducing waveform clustering. (1), waveforms from all stations recording an event are cross-correlated to group records with similar shapes and align them in a cluster stack. (2), each cluster identified in Step 1 is then evaluated by re-running the cross-correlation. A station is chosen randomly as the reference station and the time shifts relative to the reference station are recorded. (3) To maintain the same quality-control routine as used for Guy Masters previous tomography models, we obtained full travel times for the stations in each cluster by manually cross-correlating a normal-mode synthetic waveform with a “summary” waveform for each cluster. The summary waveform was simply the average of all the waveforms in that cluster aligned to the reference station based on the cross-correlation time shifts. Thus, “onset” denotes the manually identified arrival being reported, whereas “cross-correlation” denotes the method used to construct and align the cluster

and transfer the onset pick to its member records; it does not mean that the reported pick is measured by a cross-correlation fitting to a synthetic waveform, which usually aligns to the maximum amplitude instead of the onset. To clarify, we have rewritten the paragraph in Section 2 to state this two-step procedure directly. We hope the revised text is clear and does not require a schematic.

Regarding the ray-based interpretation, we have clarified in the Discussion that our argument is comparative rather than dependent on strictly ray-like sensitivity of individual measurements: the wavefront healing hypothesis predicts a specific sense of frequency dependence between the short period (~ 2 s) and long period (~ 20 s) picks, and the observed change in slope has the opposite sense. Because our measurements are onset times, a delayed onset directly records a slowdown of the earliest arriving energy in the measured frequency band rather than a distortion of the waveform shape, which is why the comparison between the two catalogs is diagnostic even if the sensitivity of each individual measurement is more distributed than a ray.

4) *What new capabilities are enabled beyond automation?*

Does the new method improve performance for low signal-to-noise records that were rejected in Houser et al. (2008), or does it mainly enable large-scale automation at similar quality thresholds? If there are additional advantages (e.g., fewer biases with emergent onsets, better robustness to source complexity, improved phase discrimination), please quantify them. Are there other adjustments that added advantages apart from automation? What would be required to extend the method to later phases (PP, SS, SKS, ScS, etc.)?

The main new capabilities beyond automation are that the model requires no synthetic waveform, reference velocity model, or source information at measurement time, and that every pick carries a prediction probability serving as a continuous per-measurement quality indicator. This makes the method applicable to records for which reliable synthetics are unavailable, and potentially to records with lower signal-to-noise ratios than interactive procedures accept; we have added this point to the Discussion. Since we have not yet performed the analysis required to quantify the low signal-to-noise performance relative to the records rejected in Houser et al. (2008), the revised text states this as a possibility under ongoing investigation rather than a demonstrated advantage. Extending the method to later phases (PP, SS, SKS, ScS, etc.) requires labeled training picks for those phases and a longer input window; as noted in the Discussion, we expect that an attention encoder trained jointly on additional phases will also improve the accuracy and recall for direct P and S waves. We will include these phases to further enhance the ray coverage in the following paper for velocity structure inversion.

5) *Outliers and quality control*

Line 181 mentions outliers. Are these outliers presumed to be automated picking errors, or could they be genuine large anomalies? I have observed variations in S-wave traveltime anomalies that are larger than 15 s along small-aperture regional arrays when S waves skirt the margins of LLVPs in the lower mantle.

We have made this explicit in the revised Results. Most of the outliers are picking errors: after checking the waveforms with the largest delays in GLoBoCat, they appear to be mispicks in which the model locks onto a later swing due to cycle skipping, and they show little geographic consistency. These records amount to only ~ 300 out of the $\sim 100,000$ picks in common with the ISC catalog ($\sim 0.3\%$), although they are conspicuous in the scatter plots. At the same time, we appreciate the reviewer’s observation that genuine S wave anomalies exceeding 15 s occur where raypaths skirt the LLVP margins; the revised text explains that the summary ray analysis requires consistency among neighboring raypaths, which preserves geographically coherent extreme anomalies while removing isolated mispicks. We have expanded Supplementary S2 with the waveform-inspection results and an improved visualization of the frequency distribution of the travel time anomalies.

6) *“Improved coverage” needs to be demonstrated in a figure.*

The manuscript repeatedly mentions “improved coverage” (e.g., line 193). It would strengthen the paper to explicitly show where mantle sampling has improved since 2008. Consider adding to Figure 3 or making a new figure that compares maps of turning points for GLoBoCat and the Houser et

al. catalog from 2008, and give regional summaries that explain what parts of the mantle are newly sampled.

We have added text to the Results pointing to the side-by-side maps comparing the geographic distribution of GLoBoCat travel time residuals with previous catalogs (Figs. S3–4) and summarizing the regions where mantle sampling has improved since 2008: raypaths turning beneath the South Atlantic, southern Africa, the Indian Ocean, and the South Pacific, previously sparsely sampled, are now densely covered for both P and S waves, while the coverage of the northern hemisphere is retained. We have also updated Fig. 2 with grouped histograms of turning point counts in the southern hemisphere, quantifying the improvement between GLoBoCat and the Houser et al. (2008) catalog.

7) *Interpreting $\delta \ln V_S / \delta \ln V_P$ and estimating R .*

Figure 3 shows coherent spatial patterns in $\delta T_S / \delta T_P$, and the increase of $\delta T_S / \delta T_P$ with distance beyond 80° is also evident. However, inferring the behaviour of R with depth requires more careful analysis than currently presented. In particular

- LLVPs rise hundreds of kilometers above the CMB. For example, P and S waves from SW Pacific events to North American stations and from South American events to eastern African stations can traverse the full vertical extent of LLVP structure beneath the Pacific and Africa.
- The crust and upper mantle are strongly heterogeneous and can contribute substantially to traveltime variations; so can centroid mislocations and origin-time errors. Chaves et al. (GJI, 2021; <https://doi.org/10.1093/gji/ggaa534>) provides a clear illustration of how strong upper-mantle contrasts beneath North America affect traveltimes and estimates of R .

We agree that inferring the behavior of R with depth requires more careful analysis than the travel time residual ratio presented here, and we have added a paragraph of caveats to the Discussion. The revised text emphasizes that epicentral distance is only a proxy for turning point depth and that the residual ratio is a path-integrated quantity rather than a direct estimate of the local $\delta \ln V_S / \delta \ln V_P$; that raypaths traverse the strongly heterogeneous crust and upper mantle, where structure as well as centroid mislocations and origin time errors contribute to the residuals (Chaves et al., 2021, now cited). Isolating the lower mantle contribution requires a formal inversion that accounts for three-dimensional structure and corrects for the crust and upper mantle, which will be presented in a follow-up paper built on this catalog.

8) *ISC vs GloBoCat differences.*

I recommend that the authors show representative waveforms for the “extreme” ISC-GloBoCat differences. In Figure 4b, many points indicate $dT_{gbc} = +5$ to $+12$ s and corresponding values of $dT_{isc} = -3$ to $+1$ s. It seems unlikely that a 10–15 s discrepancy can be attributed solely to physical dispersion from 3-D heterogeneity (wavefront healing), at least without additional quantitative support.

An alternative explanation is that high-frequency ISC picks belong to sharp onsets tied to rupture initiation, whereas long-period GloBoCat picks are sensitive to the first major asperity, analogous to differences between NEIC hypocenters and long-period CMT centroids. This possibility deserves explicit investigation (e.g., dependence on magnitude, source-time duration, depth phases, or event type), and waveform examples would be very helpful.

We agree and have added both an analysis and waveform examples. Visual inspection of representative records shows that the extreme differences ($\delta t_{GloBoCat} = +5$ to $+12$ s with near-zero ISC residuals) are dominated by mispicks in which the model locks onto a later swing of the waveform due to cycle skipping, corresponding to a delay of roughly a half-period. We also investigated the alternative explanation suggested by the reviewer: comparing GCMT centroid times against NEIC hypocenter times for the common events yields offsets of only ~ 3 seconds, too small to account for 10–15 s discrepancies, and this is now stated in the revised Discussion. Importantly, the fitted slopes are measured with an outlier-resistant stochastic regression separately for early and late arrivals, and no corresponding cluster appears

below the extrapolated early-arrival slope, so the change in slope is not an artifact of this small outlier population.

9) Wavefront healing / physical dispersion.

The discussion starting near line 264 on wavefront healing is intriguing, but substantially more work seems necessary to interpret the ISC-GloBoCat differences as physical dispersion. As written, it remains unclear what magnitude of dispersion is predicted for the relevant frequency bands and paths. It may be useful to compute AxiSEM3D synthetics for representative paths at 2 s (the shortest possible period) and 15 s periods for PREM and a 3-D model such as HSML. My expectation is that predicted physical dispersion would be smaller than 1 s and may be hard to distinguish from other frequency-dependent effects, such as the finiteness of the source-time function, focal depth errors affecting pP/sP/P interference, crustal reverberations influencing long-period P and S waveforms (e.g., Ritsema et al., GJI, 2009).

We agree that substantially more modeling is needed before the ISC–GloBoCat differences can be attributed to physical dispersion, and we have followed the reviewer’s suggestion (ii) below: the revised Discussion presents physical dispersion as a hypothesis with explicit caveats and alternative explanations. The new text states that we have not quantified the dispersion predicted for the relevant frequency bands and paths, and that doing so requires full waveform synthetics in three-dimensional models computed at both short (~ 2 s) and long (~ 15 – 20 s) periods, e.g., with AxiSEM3D for representative paths in PREM and a 3-D model such as HSML, which is beyond the scope of this catalog paper and is planned as dedicated future work. The alternative frequency-dependent effects raised by the reviewer—the finiteness of the source time function, focal depth errors affecting pP/sP/P interference, and crustal reverberations influencing long period P and S waveforms (Ritsema et al., 2009, now cited)—are listed explicitly in the revised text as explanations that cannot yet be ruled out.

I encourage the authors either to (i) add quantitative modeling to support the physical-dispersion interpretation, or (ii) present this as a hypothesis with clear caveats and alternative explanations.

Jeroen Ritsema
University of Michigan
(Mar 31, 2026)

As described above, we have adopted option (ii): the physical-dispersion interpretation is now presented as a hypothesis with clear caveats and alternative explanations, and the quantitative modeling required to test it is identified as future work. We thank Prof. Ritsema for his thorough and constructive review, including the annotated manuscript, whose suggestions we have incorporated in the revised text.

Reviewer 2

I have little to contribute to the ML part of this paper. The results seem reasonable and the final quality-controlled datasets have characteristics such as standard deviations (2.3 sec for P and 4.2 sec for S) which are similar to older datasets (typically about 1.9 sec for P and 4.0 sec for S). The errors in the older datasets ranged from .5 to about 2 seconds depending on individual measurement quality while here it seems all measurements must get the same error which is about 1.8 seconds for P and 2.4 seconds for S. It should not be a surprise that these automated picks would be less precise than the interactively measured picks but the authors have gotten close and there is still signal to be interpreted.

We thank the reviewer for this careful assessment. It is correct that the automated picks are somewhat less precise than the interactively measured picks, and that we quote a single error per phase for a given probability threshold. We note, however, that each pick in GloBoCat carries its prediction probability, which serves as a per-measurement quality indicator analogous to the individual quality levels in the older

datasets: users can raise the threshold to reduce the picking error (e.g., to 1.31 s for P and 1.64 s for S at $p = 0.9$) at the cost of data volume. We have added a sentence to the Results making this explicit.

I do have a couple of technical questions about the measurement algorithm. I note that the authors interpolated the time series to have a .25 second spacing. The GRFO SRO instrument response was originally chosen as a standard instrument response because it has a relatively impulsive first swing and manual picking can often be done to about 0.1 sec. I would therefore have chosen 0.1s for resampling the seismogram but I doubt it makes much difference, particularly as the picking errors are ended up being fairly large. The second thing is that the authors apparently use all 3 components (V,R,T) in their picking algorithm and it is not clear to me what the advantage of including the radial component is unless somehow particle motion is being used to identify S. It would seem naively to me that it is best to measure P from the less noisy verticals and S from the transverse components — unless the S phase is vertically polarized (such as SKS). Perhaps the authors could clarify what the advantage of using all 3 components is.

The choice of 0.25 second sampling is a trade-off between timing precision and computational cost (including the required training set size), because the number of model parameters grows roughly quadratically with the length of the input time series. Training at 0.1 s sampling would increase the neuron numbers (6.25x under the U-Net architecture) and thus require substantially more data and computation while contributing only a subtle improvement, given that the picking uncertainty is dominated by the 1.25 s width of the probability function used to represent arrivals in training. As for the radial component, the change in model size from dropping it would actually be minimal, because the channel number expands immediately in the first hidden layer to capture features; reducing the input to two components only removes about a third of the neurons in the first layer, roughly a few thousand parameters. We therefore retain all three components, following the original Earthquake Transformer and most models fine-tuned from it. Although not explicitly wired, the neural network may use the particle-motion information from radial component to discriminate P and S energy from noise. These rationales have been added to Section 2 of the revised manuscript (L142–149).

I do have some concerns with the comparison with the short period ISC data. First is that the ISC S data has many limitations. I believe it is still true that ISC S data are dominantly picked from short period verticals and, due to attenuation, have periods of about 4 seconds which are not well-recorded by short period instruments. Such picks are badly contaminated by S-to-P conversions beneath the stations and by SKS interference at distances beyond about 80 degrees (both issues much reduced by using transverse components). This latter fact makes the ISC S data almost useless for looking at the relative behavior of S and P at the base of the mantle where most of the interesting action occurs (e.g. your fig 4a). It also used to be that the ISC only associated arrivals with S if their arrival times came within 7.5 seconds of a model prediction (JB tables). Even though the ISC data set has been rebuilt using more modern rules, it would seem that large S residuals are still rare (judging from fig 4c) and I really doubt that much meaningful can be extracted from a comparison with it.

We agree with the reviewer’s assessment of the limitations of the ISC S data. The S-to-P conversions beneath the stations and the SKS interference beyond $\sim 80^\circ$ are plausible contaminants of short period vertical-component picks, and the limited number of common picks ($N=3,363$) does not allow us to categorize the data by source depth or distance to isolate these causes. The concern applies to the ISC comparison in Fig. 4c. We added a caveat paragraph in the main text (L323–328). To tackle the quality/consistency issues of ISC short-period data set, we plan to work on a multi-frequencies analysis using the picks derived from our ML-picked dataset in the near future.

The ISC P data are a different story but here we have other issues. The signal from earthquake mislocation rivals the signal from 3D structure for P waves so earthquake location must be considered carefully. It was not clear to me what earthquake locations were used for this work and if consistent locations were used for both long and short period data. The terms in equation (3) need to be clarified. I thought at first that the authors had just differenced the absolute arrival times but this would be dominated by physical dispersion and clearly we are considering residuals of some kind here — but

relative to what? Furthermore, for the larger earthquakes used here, the location “seen” by long period data and short period data will be different. This can be seen by plotting the mean residual per event as a function of earthquake moment for the long period data using residuals computed using “short period” locations such as ISC or EHB. This effect is a couple of seconds over the range of magnitudes used here but the majority of your data will be from the lower magnitudes so this may not contribute much to figure 4b.

The GLoBoCat has arrival picks at calendar (absolute, as we called) time and two sets of travel times calculated using different source (earthquake location-time) models. Since our earthquake catalog comes from GCMT, we first calculate model-derived (theoretical) arrival times using the GCMT locations and origin times together with PREM for all sources in the GLoBoCat table. Subsequently, for those picks which also appear in the ISC catalog, we obtain the aligned residual by applying the difference in absolute pick time between GLoBoCat and ISC to the *residual time* records of the ISC catalog, as described in Equation 3. In this construction, the influences of all corrections (including crustal structure), the raypath, the source model, and the velocity model are common to both terms (aligned with ISC). Similarly, the mislocation may contribute to the residual but always equally to both GLoBoCat and ISC in the Fig. 4b–c, ensuring that the comparison between the two catalogs rests entirely on the arrival picking. Suppose the extremely large anomalies are produced because of the abovementioned effects, we would have seen the slope bend toward 1. However, we agree that the traveltimes residual analysis is only preliminary and we are already working on a 3D inversion to account for the velocity structure.

What worries me most about fig4b is the band of strongly positive long period P residuals (greater than 7.5 seconds which is already huge) associated with small ISC P residuals. It is clearly this feature which is torquing the change in slope of the fitted line. Such large positive residuals are absent from the earlier (“semi-manual”) data sets. Are these residuals geographically concentrated (in previously un-sampled southern hemisphere regions for example) and are there correspondingly huge S wave residuals? If so, then perhaps they are believable and the question of why they are so different from the short period data becomes a pressing one. However, if I were the authors, I would be visually inspecting these particular records to see exactly what is being measured before making any such claims.

We visually inspected representative records from this band, as the reviewer recommended, referencing the GCMT-predicted polarities. The geographic distribution of the large positive P residuals shows no clear pattern (in particular, no concentration in the previously under-sampled southern hemisphere), there are no correspondingly huge S residuals with geographic coherence, and the inspected waveforms are dominated by mispicks in which the model locks onto a later swing of the waveform due to cycle skipping; these mispicks amount to $\sim 0.3\%$ of the picks in common with the ISC catalog. Importantly, the change in slope is not torqued by this band: the slopes are measured with an outlier-resistant stochastic regression separately for negative and positive ISC residuals, and the cluster that would extend the early-arrival slope into positive residuals is absent, which is what allows us to make the claim. This analysis has been added to the Discussion and Supplementary S2.

One final thing. The digital data prior to about the year 2000 had many issues with timing though these timing errors were usually less than 1 or 2 seconds (or grossly wrong by a minute or more). I spent a lot of time documenting these and correcting the final data for them. While these were often data logger issues, you will also be sensitive to errors in instrument responses (because you are re-responding your data and rotating horizontals). I would not be surprised if a significant fraction of instrument responses have issues, and these can lead to apparent timing issues. This is why procedures like your summary ray analysis are so critical, though it will not help with isolated stations. I suspect you also record polarities of arrivals and cross-checking these can also reveal problems. While I would hope that the larger networks have QC under control (though history would not support this), I am guessing that some smaller local arrays and the occasional temporary deployment may still be subject to problems. It is worthwhile keeping this in mind when looking at unusual data.

We thank the reviewer for sharing this experience. Because GLoBoCat spans 2010–2023, it avoids most of the data logger timing problems documented for the pre-2000 digital data. However, we agree that

errors in instrument responses can lead to apparent timing issues, since we “re-response” the data to the reference GRFO response and rotate the horizontals. Our quality control depends heavily on the summary ray analysis, which filters 10–15% of the data in each depth range but, as the reviewer notes, does not protect isolated stations. We did face the timing issue with the local network NZ (GeoNet) and we have to remove the measurements, as reported in Supplementary S1. We now remove measurements with problematic cycle skipping for each pick in the GLoBoCat-ISC comparison by cross-checking the first-motion polarities against the GCMT mechanisms. This is a promising additional screen that we plan to incorporate in future versions of the catalog. We will keep the reviewer’s warning about smaller local arrays and temporary deployments in mind when examining unusual data.