We thank Dr. Conevski for the detailed and constructive review, which engaged deeply with both the conceptual framing and the technical workflow of our study. In response, we restructured the benchmarking framework and revised all related scripts. We introduced a staged quality-control protocol, reprocessed both datasets, and added more detailed interpretability and uncertainty analyses. Each change is described below in response to the reviewer's specific comments, and we address the main concerns using the new evidence from the revised analysis.

Summary of principal changes

1. We introduced staged bottom-track velocity filtering protocols from BT_Vel_stage0 to BT_Vel_stage4. The previous target calculation was removed and replaced with the corrected horizontal bottom-track velocity definition. This correction raised the field cross-validated R^2 from about 0.60 to about 0.81 using the same predictors and model structure. We selected BT_Vel_stage1 as the main target for both datasets, while the other stages are used for sensitivity analysis.
2. Linear baselines, including OLS, Ridge, and Lasso regression, were added to every model comparison.
3. Relative bottom-track backscatter strength (BS_rel) was extracted from the raw acoustic files and evaluated as an additional predictor for both datasets.
4. Interpretability was expanded beyond global SHAP values. We added permutation importance, Accumulated Local Effects, local SHAP, conformalized prediction intervals, and a physical-versus-instrument feature-set comparison for both datasets.
5. The scope and novelty of the study were clarified more directly in the revised manuscript.
6. Because the target definition and filtering workflow changed, the model rankings also changed. We updated the model analysis accordingly and also tested hyperparameter tuning and manual tuning to examine the weakness of the deep-learning models.

Major Critical Point 1 - Circular logic and model validity

The authors use apparent bedload velocity (va) as the target variable while using input variables from the same instrument. The manuscript claims these models will perform same good if the target variables change, but provides no comparative evidence. Without a physical baseline or cross-instrument validation, the model risks being a "black box" with limited transferability to other target sets. The va – targets and the input feature set are coming from a same instrument. So the major question, why would someone need these models, when the variable is already available?

We agree with this concern, and to address it we have followed the reviewer's suggestion by clarifying the scope and novelty of the study in the revised manuscript. The purpose is not to replace direct bedload measurements or to claim that absolute bedload transport rates can be predicted without calibration. Instead, the aim is to examine how bottom-track velocity, used

here as a proxy for near-bed sediment activity, is related to other ADCP-derived hydraulic and acoustic variables. Furthermore, bottom-track-derived velocity is not always continuously usable, as it depends on a reliable bed return and may be reduced by quality-control filtering. In contrast, water-column velocity and acoustic variables are routinely available during ADCP measurements. The analysis therefore focuses on what ML and DL models can learn from ADCP data, which models better capture the hydraulic–proxy bedload relationship, and how quality-control filtering affects model performance. We also clarify that the model has practical value for interpretation, quality control, sensitivity testing, and reproducible benchmarking, and that the outcome can serve as a baseline for future studies where ADCP-derived proxy variables are combined with direct physical sediment transport measurements.

First, we corrected the target definition. In the old version, BT_Vel was obtained by averaging bottom-track velocity data after setting negative values to zero. This was too simplified and did not represent the horizontal apparent bed-velocity signal clearly. In the revised workflow, we defined BT_Vel_stage0 as the horizontal bottom-track velocity magnitude using the east and north components only. This correction alone increased the field cross-validated R^2 from 0.603 in the old workflow to about 0.812 using the same predictors and models. We then used BT_Vel_stage1, which applies the first-stage direction-consistency filter, as the main target for both datasets. The field performance remained almost unchanged at $R^2 = 0.808$, showing that the corrected target contains a clearer physical signal and that the result is not simply due to aggressive filtering.

Second, we found that the Random Forest model produced very similar cross-validated performance in the two datasets: $R^2 = 0.809$ for the laboratory flume dataset (stationary instrument, N = 11,119) and $R^2 = 0.808$ for the field dataset (moving platform, N = 5,230), using the same feature set and model configuration. We do not interpret this as proof of full transferability, because the two datasets still have different measurement conditions and no independent bedload measurements were available. However, the comparable performance across a controlled laboratory setup and a natural-river deployment suggests that the model is not driven only by a dataset-specific or instrument-internal artefact. Rather, it indicates that part of the learned relationship may reflect consistent links between water-column hydraulic/acoustic variables and the corrected bottom-track velocity. (Figure R1)

Third, we tested whether the model skill mainly comes from physically meaningful hydraulic information or from ADCP configuration-related variables. The main benchmarking model used seven ADCP-derived predictors, but for this sensitivity test we added a few additional instrument-setting parameters so that the two feature groups could be evaluated more clearly. The physical/hydraulic group included mean velocity, depth, velocity standard deviation, correlation, expected velocity standard deviation, and SNR. The ADCP-settings group included bin distance, cell size, blanking distance, and error velocity. When tested separately, the physical feature set alone produced strong results in both datasets, with $R^2 = 0.77$ in the laboratory and $R^2 = 0.78$ in the field. These values were higher than the settings-only feature set, which gave $R^2 = 0.67$ in the laboratory and $R^2 = 0.74$ in the field. The combined feature set performed best in both cases. This suggests that the model performance is not mainly controlled by instrument settings, although these settings still add useful information when combined with

the physical variables. In this sensitivity test, the combined feature set produced the highest R^2, indicating that selected ADCP-setting variables can complement the hydraulic predictors when they carry relevant measurement information. (Figure R2)

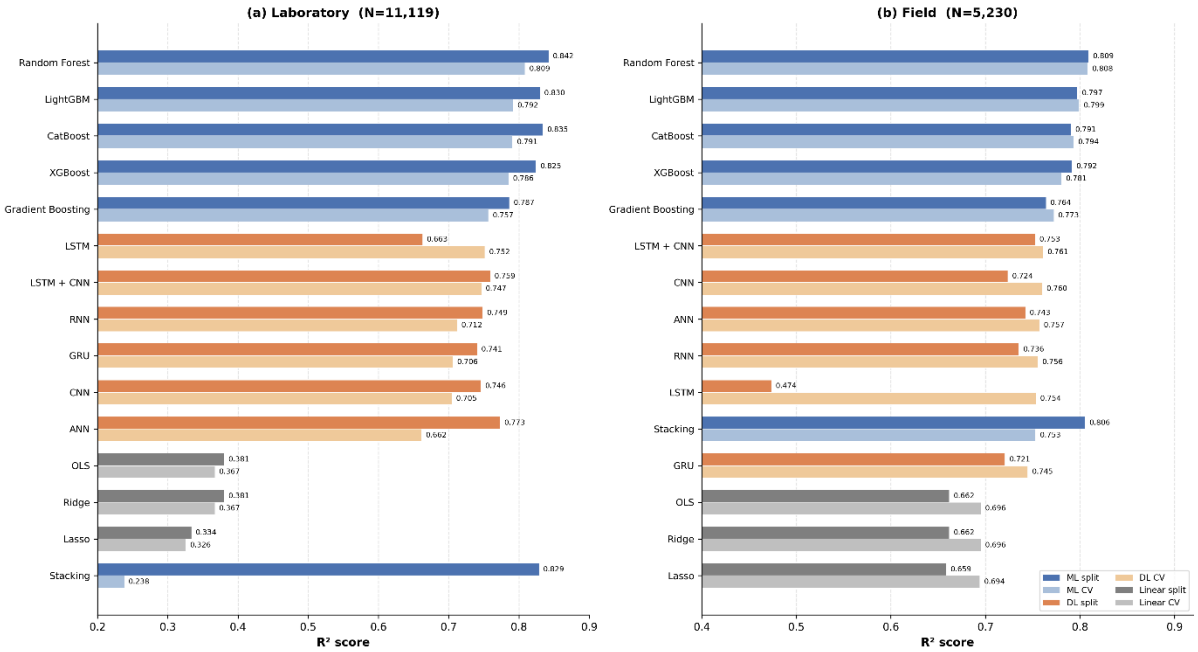

Figure R1. Cross-dataset model ranking on BT_Vel_stage1. Random Forest reaches R^2 = 0.809 (lab, N = 11,119) and 0.808 (field, N = 5,230) with identical predictors. Linear baselines shown in grey.

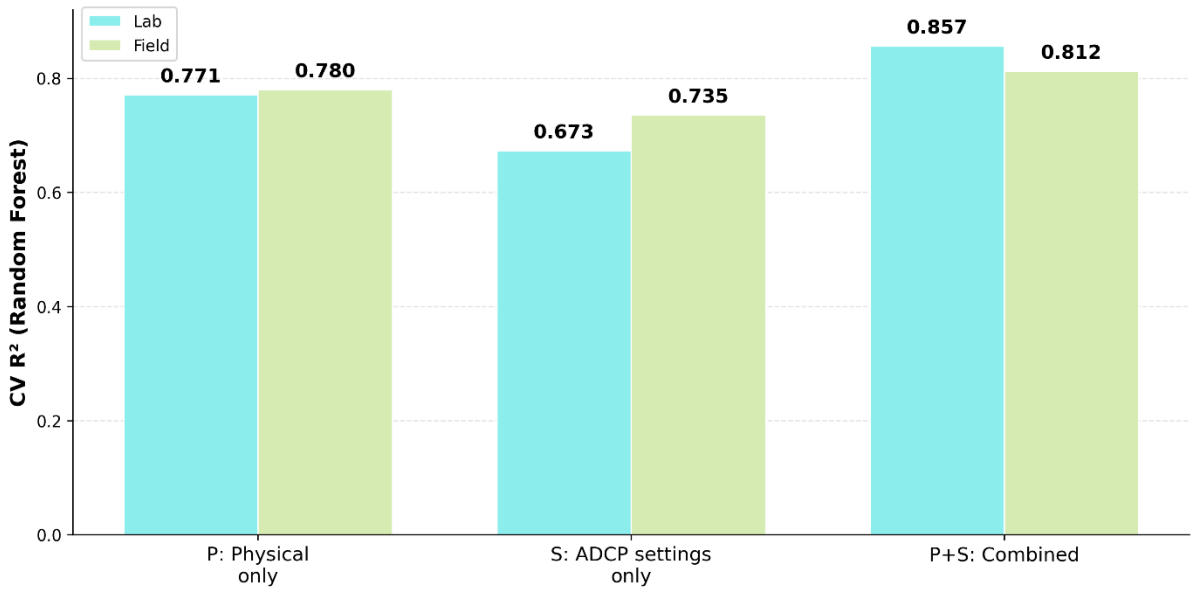

Figure R2. Physical versus ADCP-settings feature groups (disjoint). Physical features dominate on both datasets; the combination performs best.

Overall, we have revised this section to address the circularity concern more carefully. The corrected target definition, staged filtering, and feature-set comparison help clarify that the analysis is intended as an interpretable ADCP-based proxy framework, not as a replacement for direct sediment measurements.

Major Critical Point 2 - Missing Quality Control:

The filtering protocols established in Conevski et al. (2019, 2020 - JHR) for va are notably absent. Given that ADCP data is prone to noise, bypassing these standard filtering steps undermines the reliability of the training data.

We agree with the reviewer that the original manuscript did not include a sufficient quality-control procedure for the bottom-track velocity target. In response, we introduced a staged QC-sensitivity workflow guided by apparent bedload-velocity filtering concepts used in previous ADCP bedload studies. The revised workflow first defines the corrected horizontal bottom-track velocity magnitude from the east and north bottom-track components (BT_Vel_stage0). We then apply a direction-consistency filter (BT_Vel_stage1), which is the step most closely aligned with streamwise-direction filtering. Quality control in lab stage0→stage1 *raised* R^2 from 0.699 to 0.809 despite removing half the data, suggesting that the direction filter reduced noise while retaining useful signal. Additional sensitivity stages were then added: BT_Vel_stage2 combines the direction filter with an error-velocity filter, BT_Vel_stage3 combines the direction filter with a magnitude/outlier filter, and BT_Vel_stage4 combines the direction, error-velocity, and magnitude filters into the strictest QC target. We selected BT_Vel_stage1 as the main target because it provides a physically interpretable first-stage filter while retaining sufficient samples in both datasets. The later stages, especially BT_Vel_stage3 and BT_Vel_stage4, are reported as stricter QC-sensitivity tests rather than as progressively more certain ground truth. We are grateful to the reviewer for directing us to apply these established standards, which materially improved the reliability of the training data.

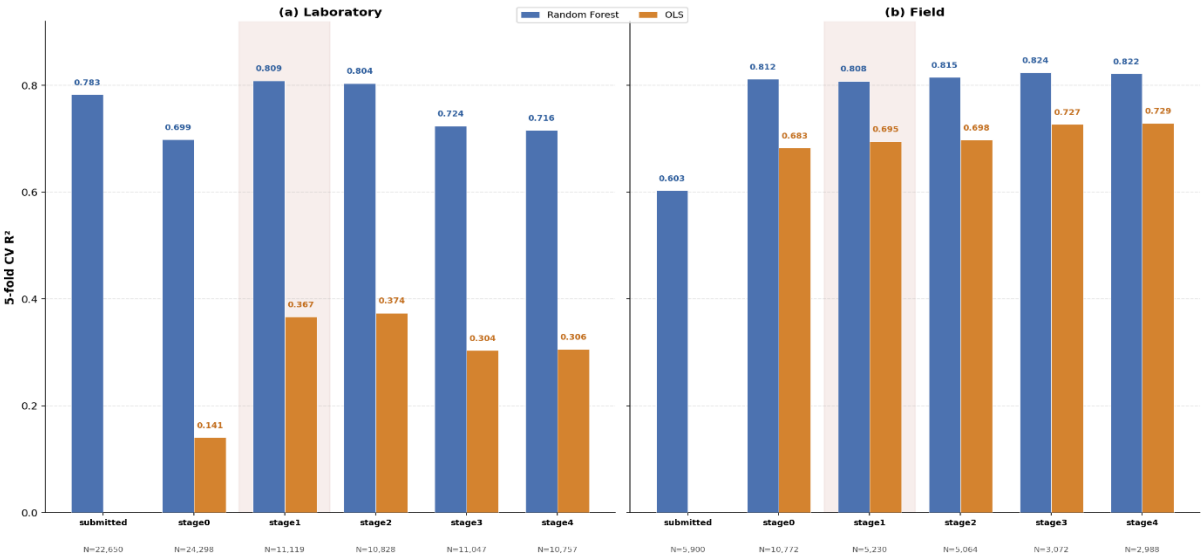

Figure R3. Filtering sensitivity across QC stages (RF and OLS). Stage 1 (shaded) is the adopted target.

Major Critical Point 3 - Missing backscatter

The study does not account for riverbed Backscatter strength (BS). As BS is a primary indicator of sediment concentration and bed properties, its exclusion from a bedload estimation model is a significant gap.

We thank the reviewer for pointing out this important gap. We agree that bottom-track backscatter strength is physically relevant for sediment-transport proxy analysis and that its absence from the original manuscript was a limitation. In the revised workflow, we extracted bottom-track echo intensity (Bt.Strength) from the raw RS5 .rsqmb acoustic files, which is not included in the standard MATLAB export. We then computed a relative backscatter index (BS_rel) using a range- and absorption-corrected formulation guided by the ADCP backscatter literature and the approach used in previous bedload studies

$$BS_{rel} = EI_{BT}+20\log10(R_{BT})+2\alpha R_{BT}$$

where EI_{BT} is the mean bottom-track echo intensity, R_{BT} is the mean range to the bed, and α is the freshwater absorption coefficient. We treat BS_rel as a relative acoustic proxy rather than an absolute calibrated backscatter strength, because the full riverbed backscatter formulation requires instrument-specific calibration terms, including source-level and transmit-power corrections, which were not available from the RS5 output. This is now stated as a limitation, and we acknowledge that applying the full calibrated backscatter equation may further improve the acoustic characterization in future work.

We tested BS_rel as an additional predictor alongside the original seven-feature set. In the laboratory dataset, adding BS_rel consistently improved cross-validated R^2 across the tested models, with gains up to +0.029. This suggests that the bed echo signal adds useful information under controlled flume conditions. In the field dataset, the improvement was smaller and model-dependent, with gains up to +0.011, which is reasonable because natural-river backscatter can be influenced by suspended sediment, bed roughness, range effects, and changing hydraulic conditions. Therefore, we added BS_rel to the revised analysis and discuss its dataset-dependent contribution as part of the sensitivity and interpretability results.

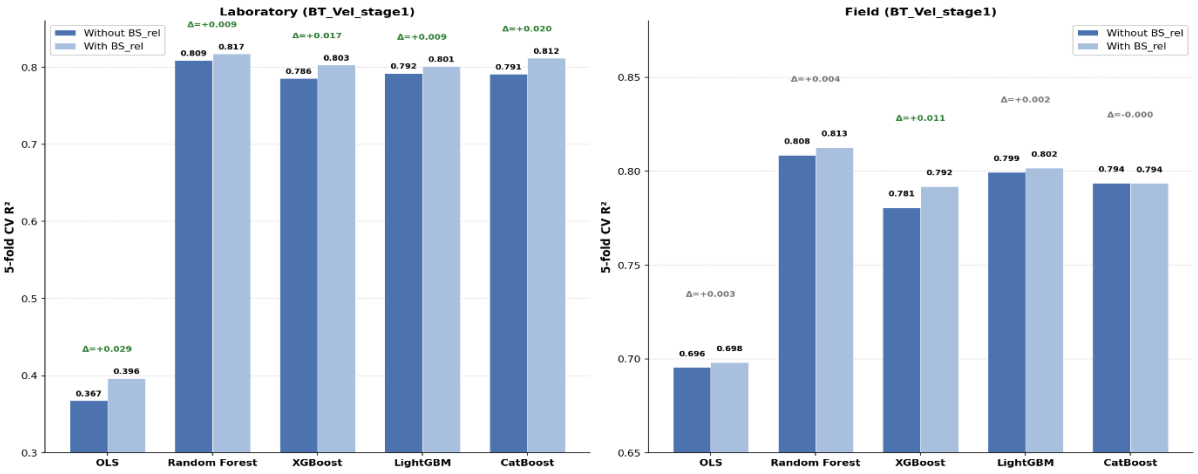

Figure R4. Effect of adding BS_rel on stage1 R^2. Meaningful gains in the laboratory (up to +0.029); smaller and model-dependent gains in the field.

Major Critical Point 4 — Interpretability

Missing more detailed interpretability analysis.

SHAP feature-importance analysis:

We agree that the original manuscript did not provide enough interpretability analysis. We added revised Random Forest SHAP feature-importance analysis for both the laboratory and field datasets. SHAP was used to identify which predictors contributed most strongly to the model output, based on the mean absolute SHAP value. This provides a clearer view of the main variables controlling the corrected BT_Vel_stage1 target. In the laboratory dataset, mean speed, depth, and correlation were the dominant predictors, while in the field dataset, depth, mean speed, and bin distance were most important. These results are consistent with the idea that the model response is mainly linked to hydraulic and acoustic measurement conditions rather than only to a black-box statistical pattern. We present this SHAP analysis as the first layer of interpretation, complemented by the additional methods described below.

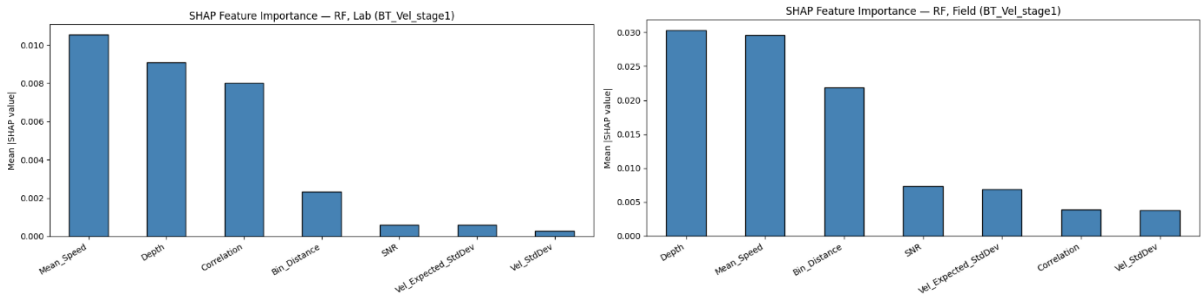

Figure R5. Global SHAP

SHAP Beeswarm:

We also added SHAP beeswarm plots to show not only the ranking of important features, but also the direction and spread of their effects on individual predictions. In the laboratory dataset, the beeswarm plot shows that mean speed, depth, and correlation have the strongest and most variable influence on the Random Forest predictions. In the field dataset, depth and mean speed dominate, while bin distance also shows a clear contribution. This additional view helps explain how the same predictors can affect the model differently across the controlled flume and natural-river datasets, rather than only reporting a single global importance ranking.

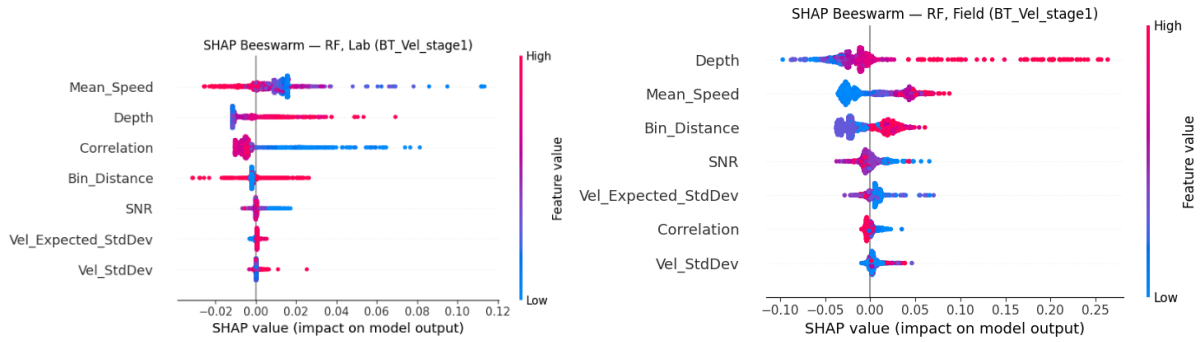

Figure R6. SHAP beeswarm

Permutation-importance analysis:

Permutation-importance was included to check whether the important variables identified by Random Forest SHAP were also relevant for other model types, rather than being specific to one algorithm. Permutation importance measures how much model performance decreases when one predictor is randomly shuffled, so larger drops indicate stronger model dependence on that predictor. In the laboratory dataset, mean speed, depth, and correlation were consistently important across models. In the field dataset, mean speed and depth again showed the strongest cross-model relevance, followed by bin distance. We note that tree-based models distribute importance between depth and mean speed, while linear models weight mean speed more heavily; we report both rather than forcing a single ranking. This supports the SHAP results and gives a more model-independent view of feature relevance.

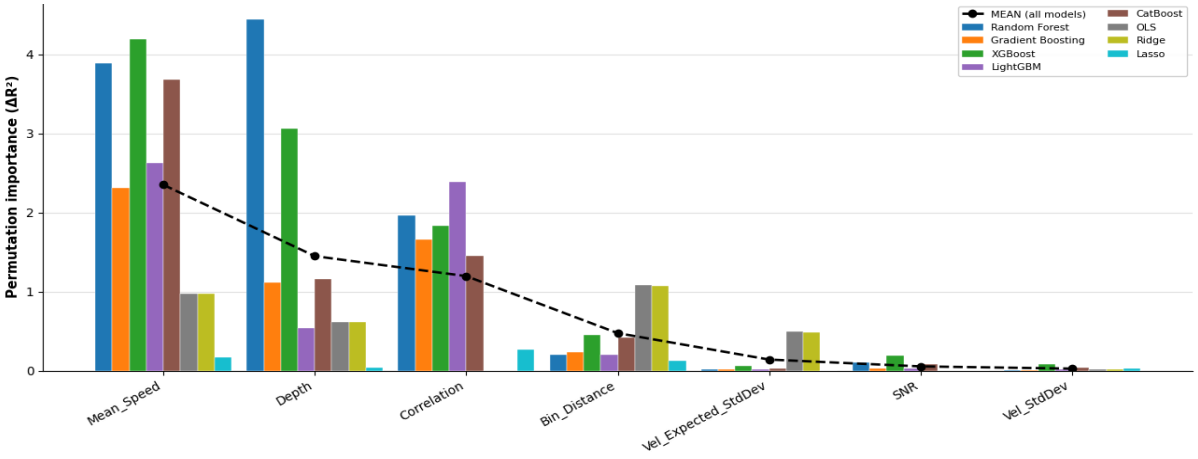

Figure R7. Permutation importance across all eight models for lab dataset. Mean speed, depth, and correlation are consistently important.

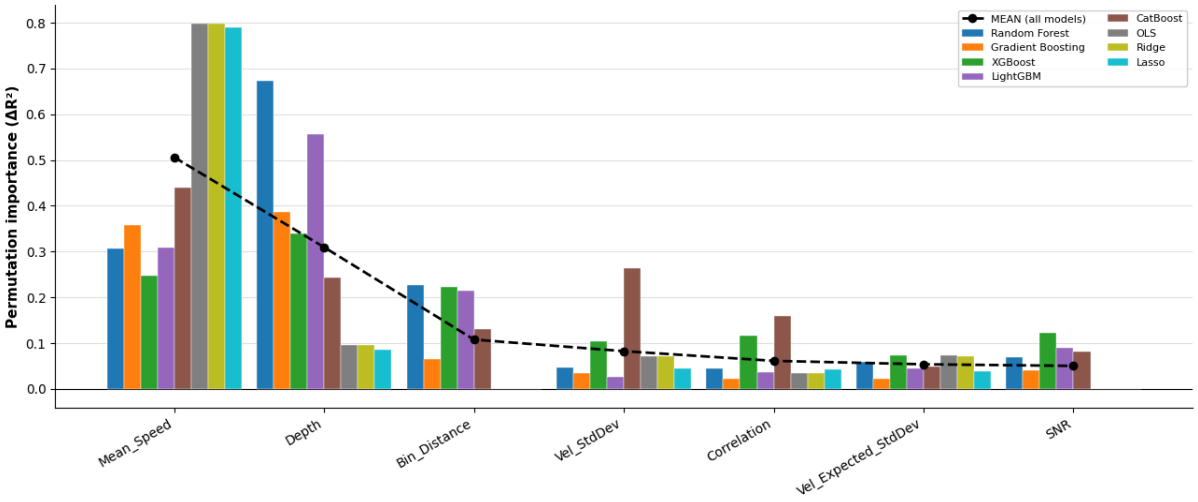

Figure R8. Permutation importance across all eight models for field dataset. Hydraulic predictors dominate; trees and linear models differ on the ranking of depth vs mean speed.

Accumulated Local Effects (ALE)

Accumulated Local Effects (ALE) plots to examine how important predictors influence the Random Forest predictions across their observed value ranges. This was included because global importance scores identify which variables matter, but they do not show the shape or direction of the model response. The ALE plots show several nonlinear effects. For example, depth has a strong positive effect in the field dataset, while mean speed shows a more peaked response, increasing over part of the range and decreasing at higher standardized values. In the laboratory dataset, depth, mean speed, and correlation show clear nonlinear patterns. These results help explain the model behavior in a more physically interpretable way, beyond simply ranking the predictors.

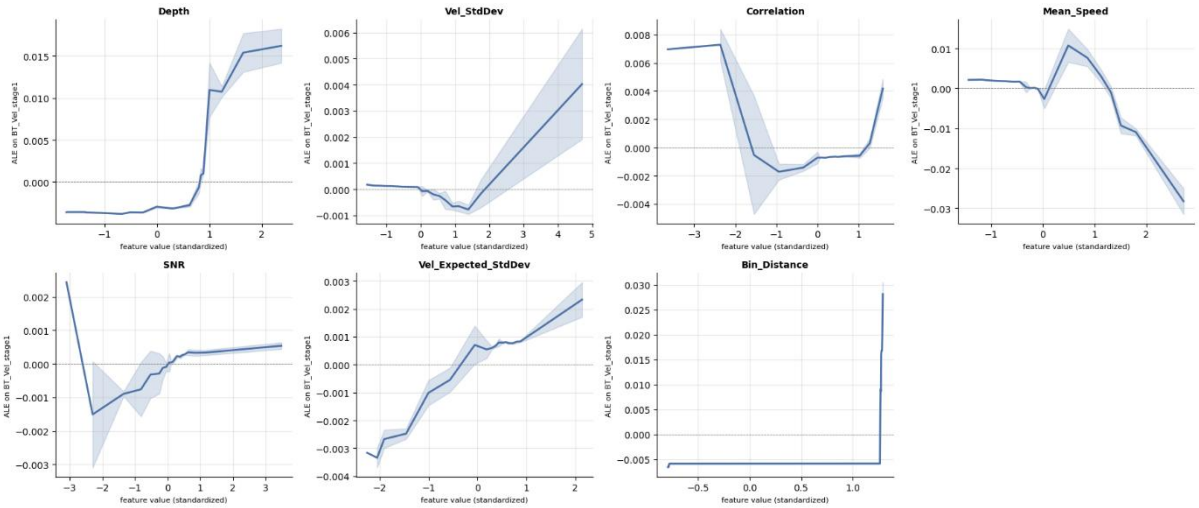

Figure R9. ALE plots - Random Forest, lab dataset. Nonlinear effects for depth, mean speed, and correlation.

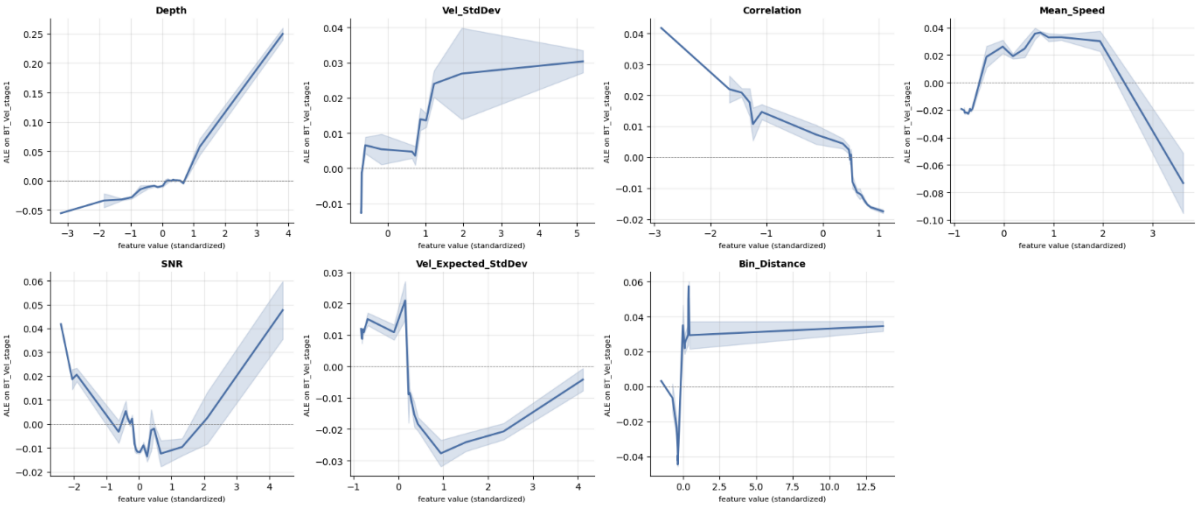

Figure R10. ALE plots - Random Forest, field dataset.

Local SHAP:

We added local SHAP waterfall explanations for selected high- and low-BT_Vel_stage1 predictions in both datasets. These plots show how individual predictors move a single prediction from the model baseline to the final predicted value. In the high-velocity examples, depth and mean speed provide the strongest positive contributions, especially in the field dataset. In the low-velocity examples, low depth, low mean speed, or lower bin-distance conditions reduce the predicted bottom-track velocity proxy. This local analysis complements the global SHAP and ALE results by showing that individual predictions are driven by physically interpretable combinations of hydraulic and acoustic variables, rather than only by an overall feature ranking.

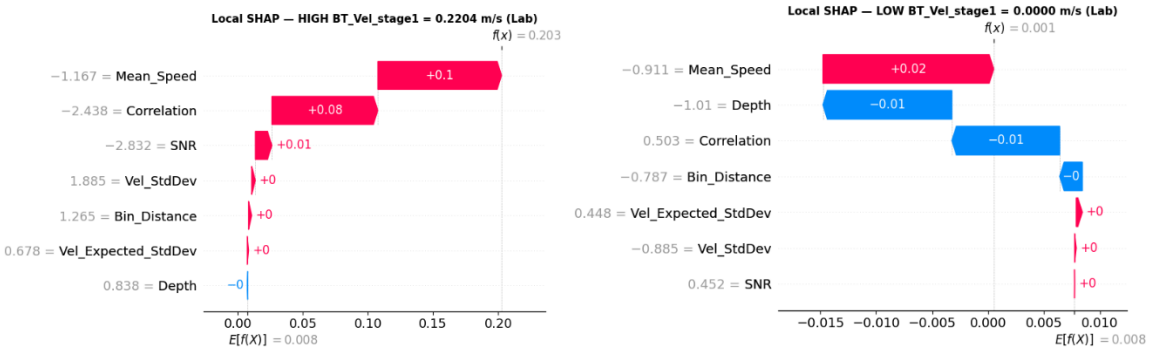

Figure R11. Local SHAP – lab; a) High BT_Vel_stage1 b) Low BT_Vel_stage1

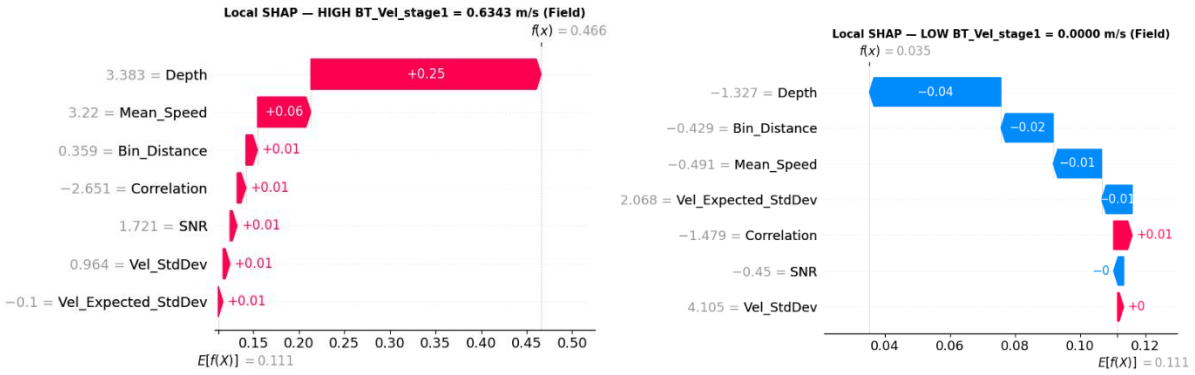

Figure R12. Local SHAP – Field ; a) High BT_Vel_stage1 b) Low BT_Vel_stage1

Prediction Intervals:

Finally, we added split-conformal prediction intervals to quantify the uncertainty of individual Random Forest predictions. This was included because model interpretability alone does not show the expected uncertainty range around the model output. The conformalized intervals achieved the intended 90% coverage in both datasets, with 90.0% empirical coverage in the laboratory dataset and 90.2% in the field dataset. The mean interval width was much smaller in the laboratory case, 0.0116 m/s, compared with 0.1174 m/s in the field case. This difference is expected because the field dataset contains more variable hydraulic, acoustic, and bed-condition effects than the controlled flume dataset. We therefore use the prediction intervals to

show not only the model estimate, but also the uncertainty range associated with each prediction.

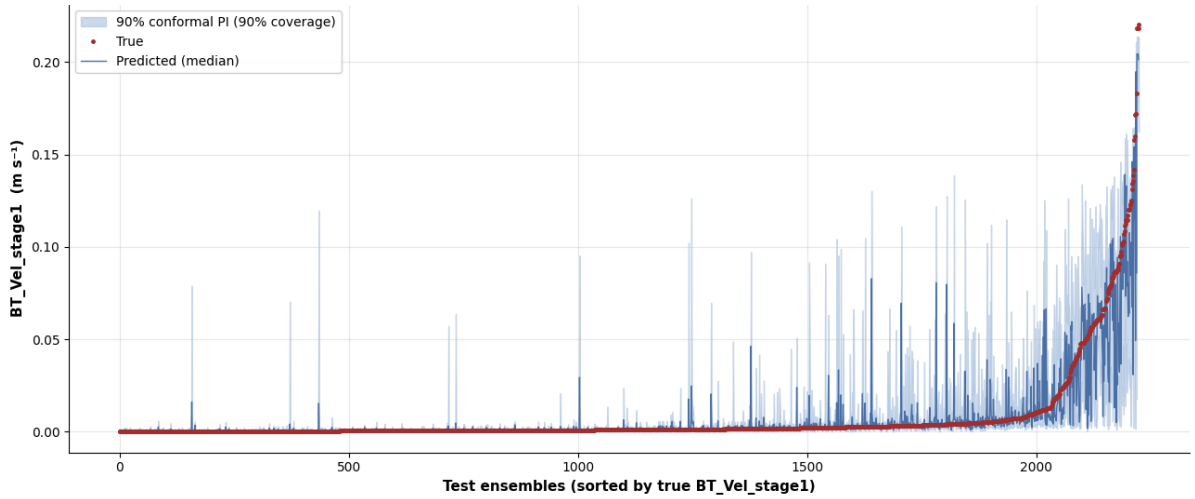

Figure R13. Split-conformal prediction intervals; laboratory (90.0% coverage, width 0.0116 m s^{-1}).

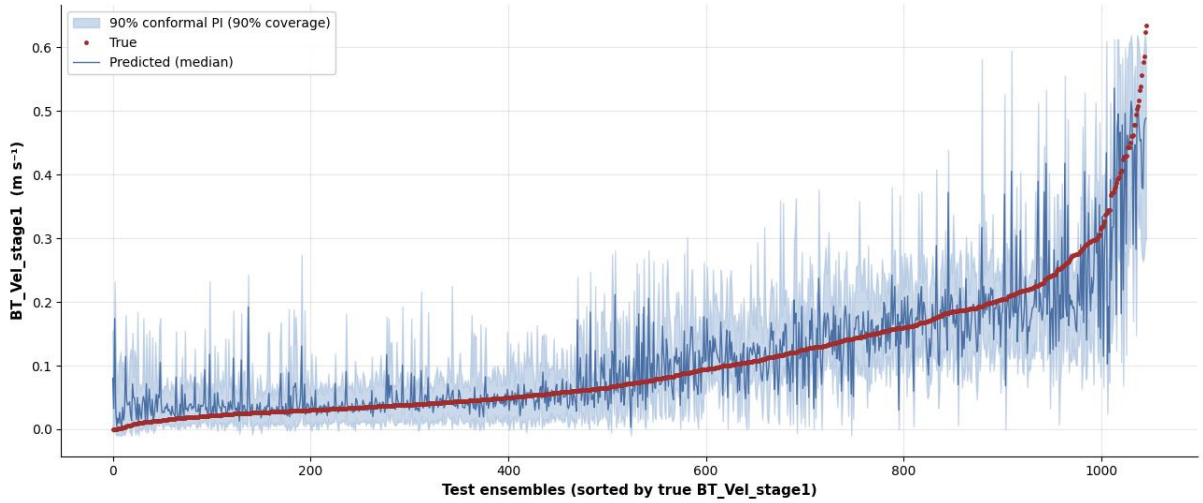

Figure R14. Split-conformal prediction intervals; field (90.2% coverage, width 0.1174 m s^{-1}).

Suggestions for Improvement:

- *Refocus the Scope/Novelty: The study would be more impactful if it shifted focus to the relationship between Bottom Track (BT) variables (va, BS) and water profiling data (the rest of the input features). This would allow for a detailed analysis of how local flow hydraulics relate to bedload transport—a topic currently under-researched….*
- *Sensitivity to Filtering: Use va at different stages of processing as target variables (e.g., va1 with only direction filtering vs va2 with the full 4-step protocol, run all the ML models or the only the best ones). This would demonstrate how ML models handle raw vs. refined acoustic data.*
- *Baseline Comparison: Include simple linear regression models as a baseline. ML complexity must be justified by showing significant performance gains over traditional statistical methods.*
- *Enhanced Interpretability: Expand the ML analysis beyond global SHAP values. The authors should include Accumulated Local Effect (ALE) plots, Permutation Importance, and local SHAP explanations. Additionally, incorporating uncertainty analysis (e.g., prediction intervals) is essential for any model intended for field application.*
- *Test of different feature sets, that have physical meaning vs other more ADCP settings related (bin size, blank distance, error velocity, etc)*

We thank the reviewer for these helpful suggestions. We have addressed each of them and responses above.

- For the suggested refocusing of the scope and novelty, we have revised the framing of the study to focus more clearly on the relationship between bottom-track variables and water-column hydraulic/acoustic predictors. This is addressed in our response to Major Critical Point 1.
- For the filtering sensitivity, we reprocessed both datasets using the staged bottom-track velocity targets from BT_Vel_stage0 to BT_Vel_stage4. The results are reported as a sensitivity analysis, and the main target selection is explained in our response to Major Critical Point 2.
- For the baseline comparison, we added OLS, Ridge, and Lasso regression to the full ML/DL benchmarking framework. These linear baselines are included in the revised model comparison results and shown in Figure R1.
- For the enhanced interpretability request, we expanded the analysis beyond global SHAP. The revised manuscript now includes SHAP feature importance, SHAP beeswarm plots, permutation importance, Accumulated Local Effects, local SHAP explanations, and split-conformal prediction intervals. This is addressed in Major Critical Point 4.
- For the feature-set comparison, we added a separate sensitivity test comparing physically meaningful hydraulic/acoustic features with ADCP settings-related

variables, including bin distance, cell size, blanking distance, and error velocity. This analysis is discussed in Major Critical Point 1 and summarized in Figure R2.

Closing note:

We have completed the additional analyses and revised the supporting scripts and figures. Since the corrected target definition, staged filtering workflow, and expanded interpretability and uncertainty analyses affect several parts of the study, the manuscript will need to be revised consistently across the abstract, methods, results, discussion, and conclusions. If the editor and reviewer consider this revised scientific approach sufficient to address the main concerns, we will then prepare the clean revised manuscript and the tracked-changes version for formal resubmission.

We thank the reviewer again for the careful and constructive review. The comments have helped us improve the validity, clarity, and overall direction of the study.