Supporting Information for "PickBlue: Seismic phase picking for ocean bottom seismometers with deep learning"

Bornstein, T.[1,2], Lange, D[1], Münchmeyer, J.[2,3], Woollam, J.[4], Rietbrock., A[4], Barcheck, G.[5], Grevemeyer, I.[1], Tilmann, F.[2,6]

[1] GEOMAR Helmholtz Centre for Ocean Research Kiel, Wischhofstr. 1-3, 24148 Kiel, Germany. [2] GFZ, German Research Centre for Geosciences, Telegrafenberg, D-14473, Potsdam, Germany. [3] Institute für Informatik, Humboldt-Universität zu Berlin, Berlin, Germany. [4] Institute (GPI), Karlsruhe Institute of Technology, Karlsruhe, Germany. [5] Department of Earth and Atmospheric Sciences, Cornell University, 112 Hollister Drive, USA. [6] Institute for Geological Sciences, Freie Universität Berlin, Berlin, Germany.

November 16, 2023

List of Tables

List of Figures

#channels	channels	AACSE	ALA	ALB	ALBORAN2009	BLANCO	CAYMAN	GERSHWIN	GORRINGE	HORSESHOE	IQUIQUE	LOGATCHEV	SEACAUSE	SPOC	SUMA	TIPTEQ	Total OBS
4	Z 1 2 H	15,332	158	172	0	1,878	1,919	0	253	132	4,365	0	7,922	169	2,210	3,909	38,419
3	Z 1 2 -	34,537	61	0	0	0	83	2	85	0	192	0	234	5	0	455	35,654
	Z 1 - H	2,417	58	51	0	995	246	38	42	484	1,692	0	2,372	10	258	1,290	9,953
	- 1 2 H	0	0	27	0	0	0	0	11	0	5	0	1,485	105	0	0	1,633
2	- 1 - H	41	0	19	358	0	30	0	2	1	62	2,160	780	8	0	383	3,844
	Z - - H	317	0	5	0	9	9	16	1	37	168	0	96	0	0	151	809
	Z 1 - -	1	38	0	0	0	11	2	10	2	395	0	45	0	0	128	632
	- 1 2 -	0	0	0	0	0	0	0	0	0	0	0	30	1	0	0	31
1	- - - H	0	0	0	1,894	0	4	765	0	40	13	9,267	3,200	1,041	0	1,963	18,187
	Z - - -	0	0	0	0	0	0	3	0	0	23	0	0	0	0	1	27
	- 1 - -	0	0	0	0	0	0	8	0	0	0	0	13	0	0	0	21
	SUM	52,645	315	274	2,252	2,882	2,302	834	404	696	6,915	11,427	16,177	1,339	2,468	8,280	109,210

Table S1: Number of traces with different combinations of channels per dataset. Z indicates the vertical component and channels 1 and 2 are the horizontal components (with unknown orientations). H denotes the hydrophone channel.

True Phase Label	Prediction			
	P_{EQT}	S_{EQT}	P_{PN}	S_{PN}
P	4298	16	4315	20
S	6	5001	11	4939
Misclassification Rate	0.14%	0.32%	0.25%	0.4%

Table S2: Number of correctly and falsely classified P and S picks. Roman letter denotes the target (true) phase label. Italic letter stands for predicted phase. EQT: EQTransformer, PN: PhaseNet. The misclassification rate equals $\frac{M}{M+C}$, where M denotes misclassified and C correctly classified picks of the corresponding column. If the pick is closer to the wrong label than the wanted pick, we consider it as being misclassified. Note that PhaseNet classifies more picks correctly than EQTransformer which is due to the fact that only picks with a confidence $\geq$ 10% are considered for both pickers while the absolute amount of picks differs. With regard to the misclassification rate, PhaseNet performs much poorlier than EQTransformer for both classification tasks.

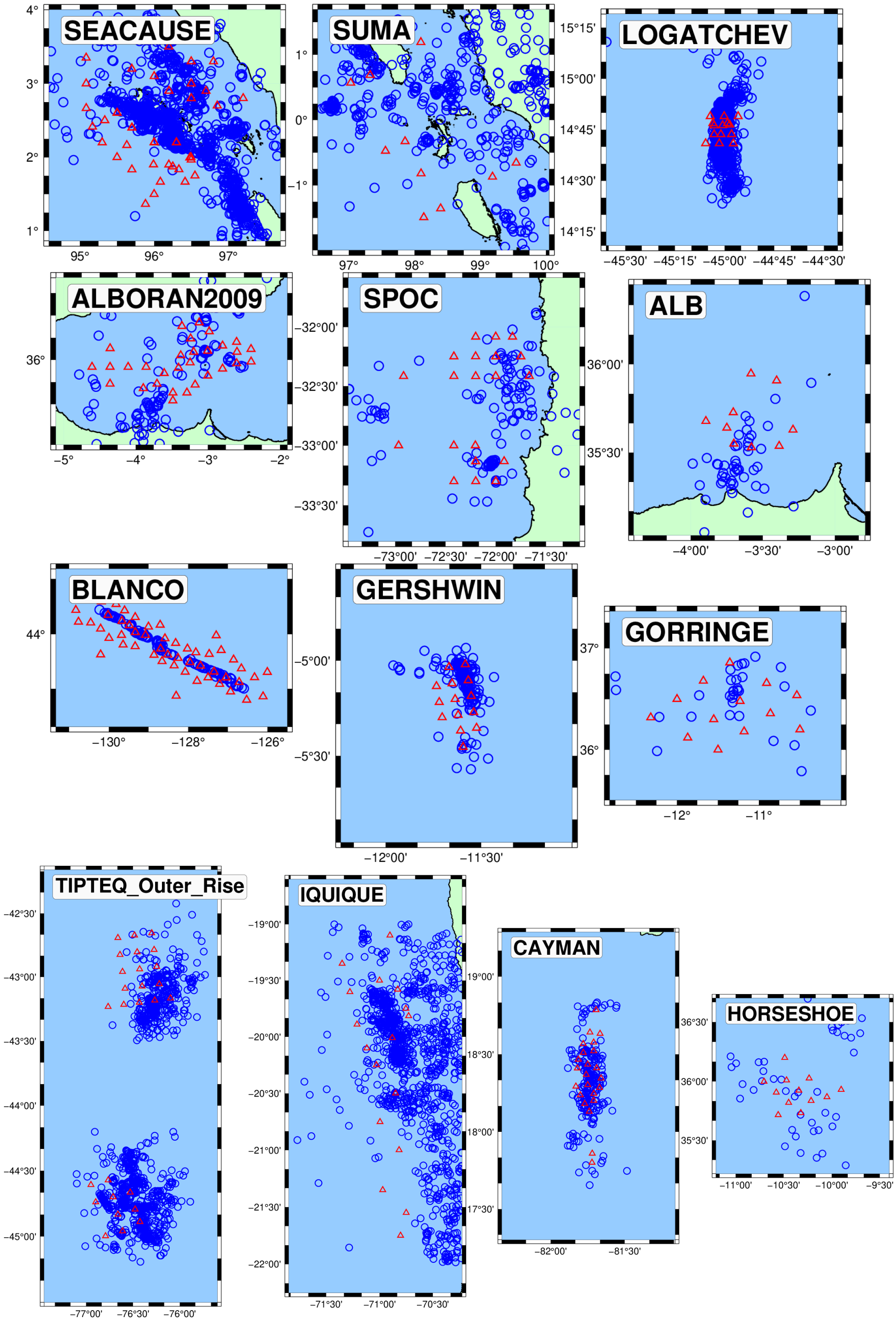


Figure S1: Close-up views for the OBS networks used in this study. Blue circles indicate events and red triangles indicate OBS stations.

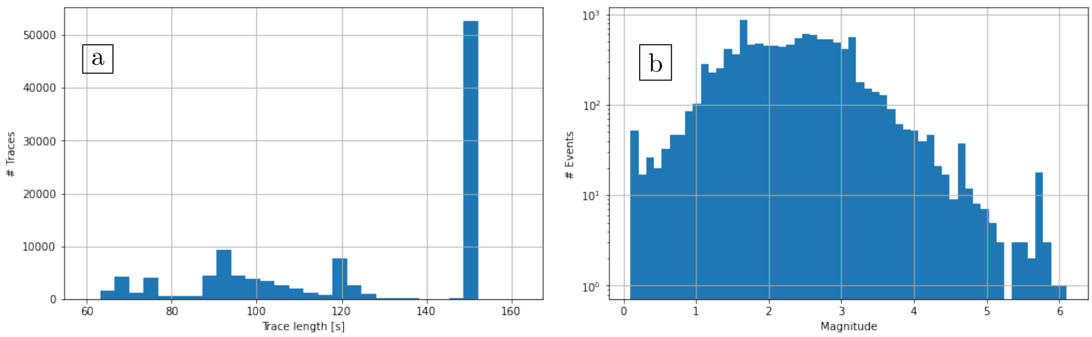


Figure S2: Trace length and magnitude distribution. Panel a: Histogram showing the trace lengths for the complete dataset. The peak at 150 s results as this is the window length for the AACSE deployment, the largest in the collection. Panel b: Histogram showing the distribution of the event magnitudes (For 8 % of the events, we do not have magnitude estimations.)

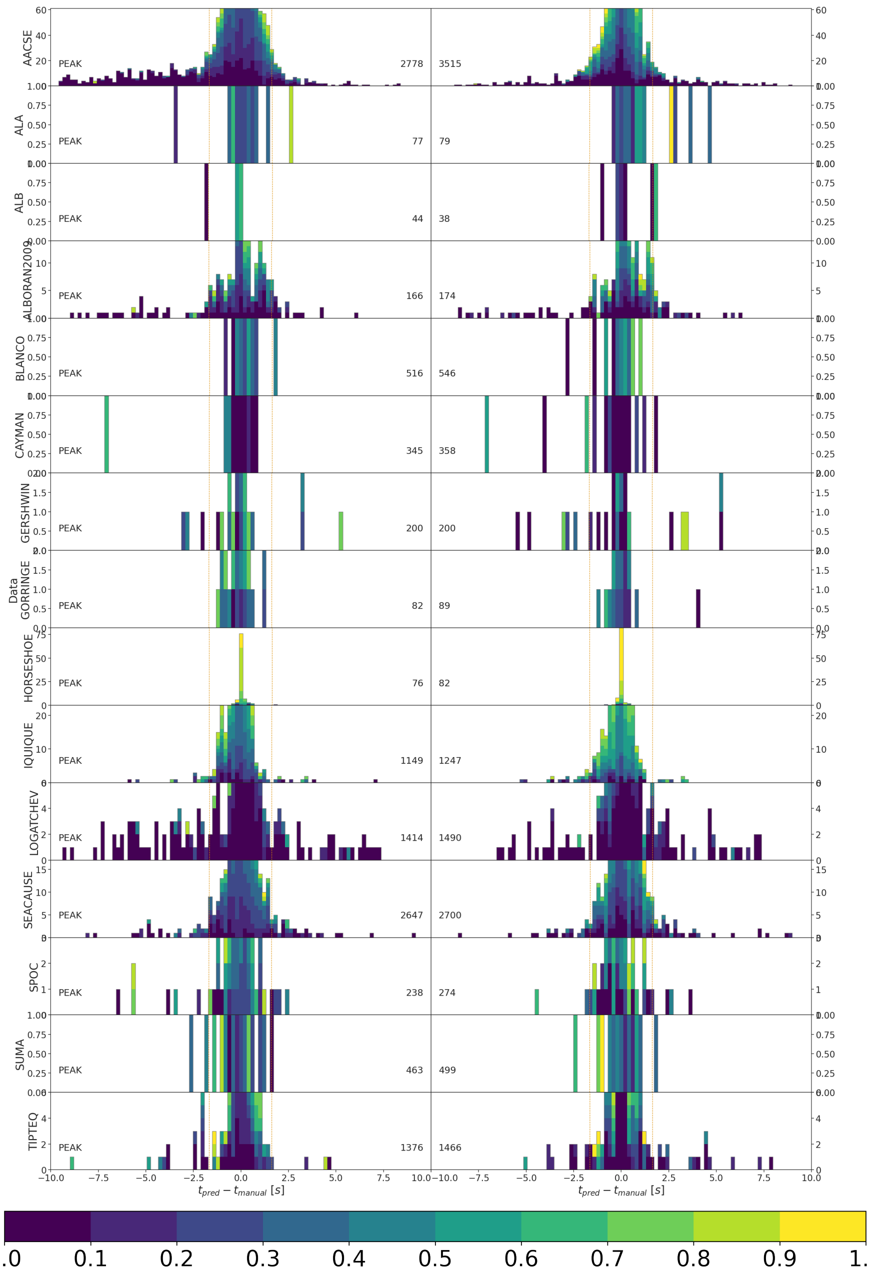


Figure S3: P residuals for individual experiments for BlueEQTransformers (left panel) and BluePhaseNet (right panel) pickers,, showing full range. 90% confidence intervals are marked with the yellow vertical lines. The models used for prediction were pre-trained on INSTANCE and then trained on the OBS dataset. The distribution is the same as that shown in Fig. 5 but with the range ±10 s.

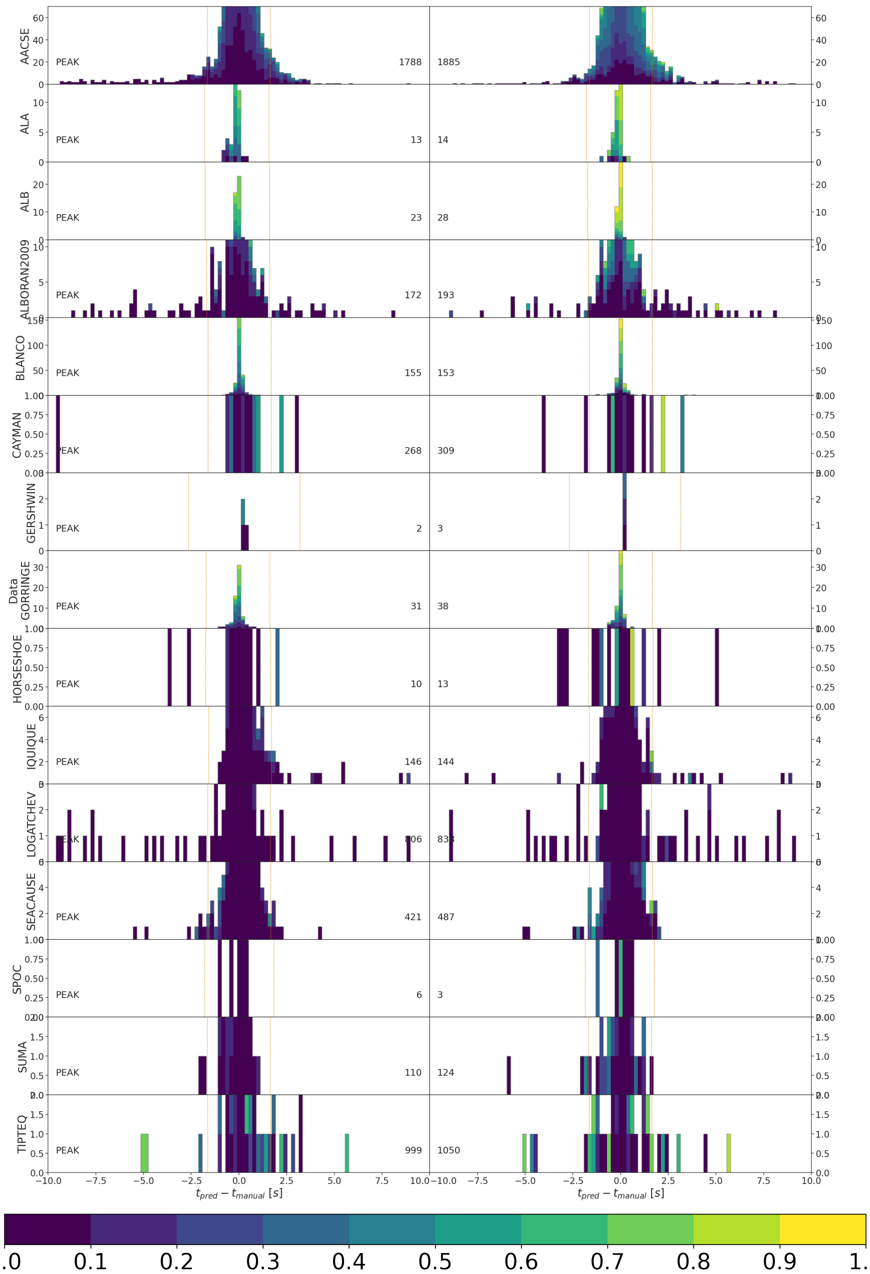


Figure S4: S residuals for individual experiments for BlueEQTransformers (left panel) and BluePhaseNet (right panel) pickers, showing full range. 90% confidence intervals are marked with the yellow vertical lines. The models used for prediction were pre-trained on INSTANCE and then trained on the OBS dataset. The distribution is the same as that shown in Fig. 6 but with the range ±10 s.

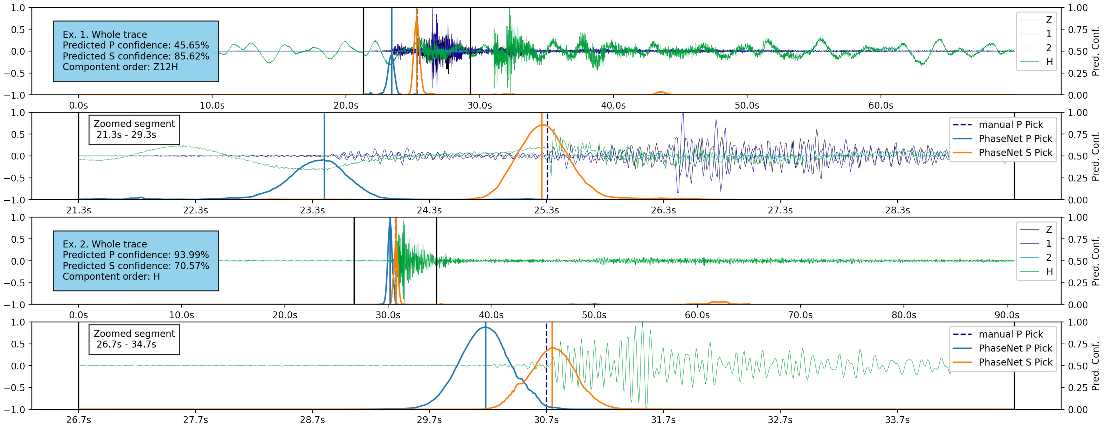


Figure S5: Waveform examples of mis-classified BluePhaseNet S picks close to manual P pick. For each example there are two plots: The first one shows the full trace while the second one zooms into the range that is marked out by the black bars in the first plot. Solid vertical lines mark predictions, while dashed lines mark the ground truth.

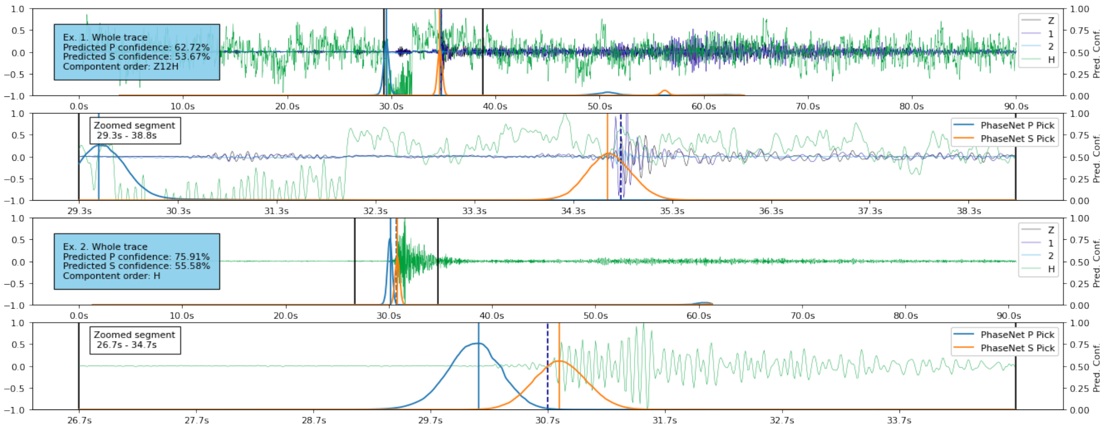


Figure S6: Waveform examples of mis-classified BlueEQTransformer S picks.

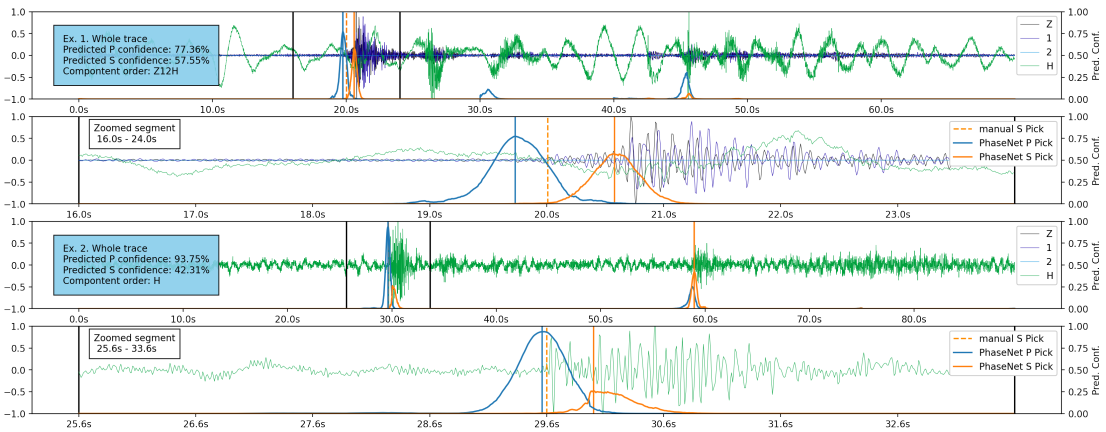


Figure S7: Waveform examples of mis-classified BluePhaseNet P picks.

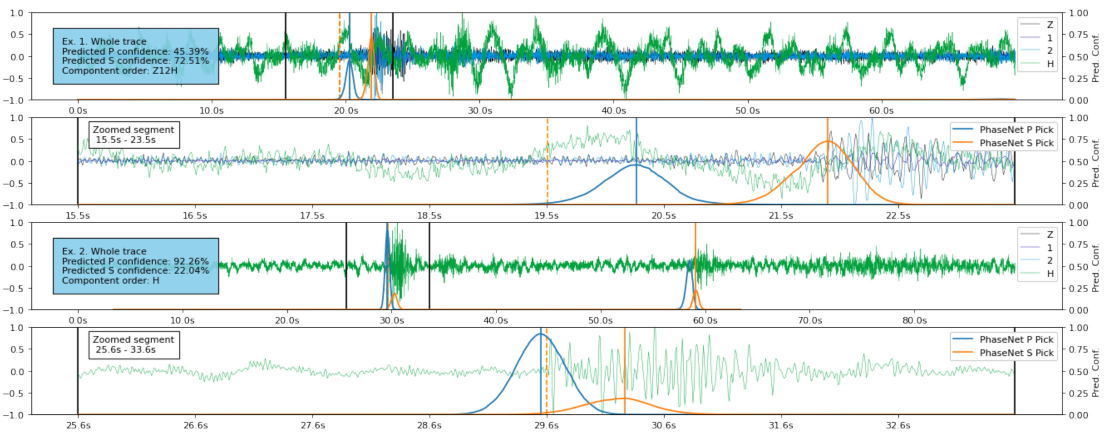


Figure S8: Waveform examples of mis-classified BlueEQTransformer P picks.

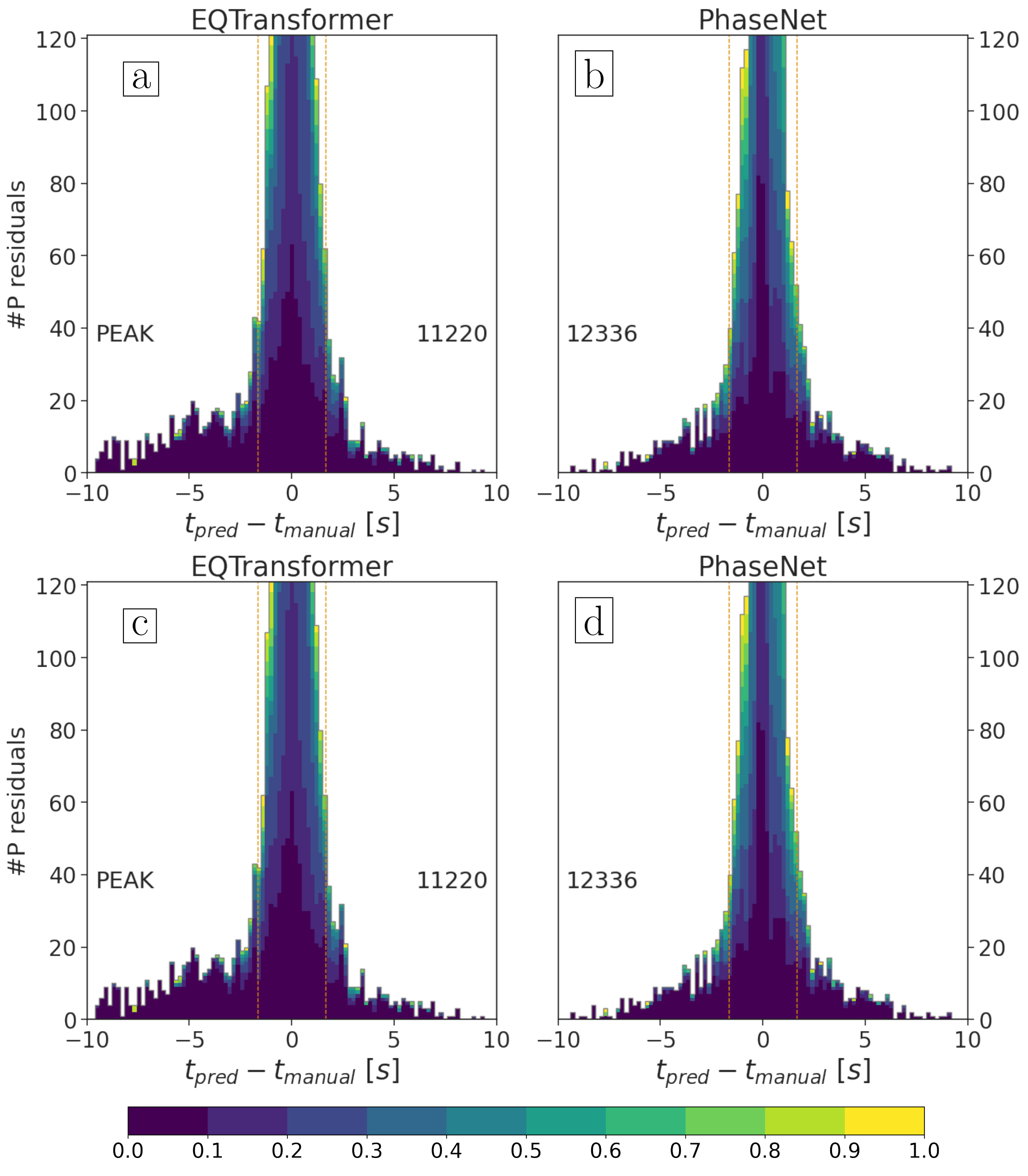


Figure S9: P residuals, pre-trained on STEAD (a,b) and without pre-training (c,d). As Fig. 9 but showing the full range to ±10 s in order to highlight outliers.

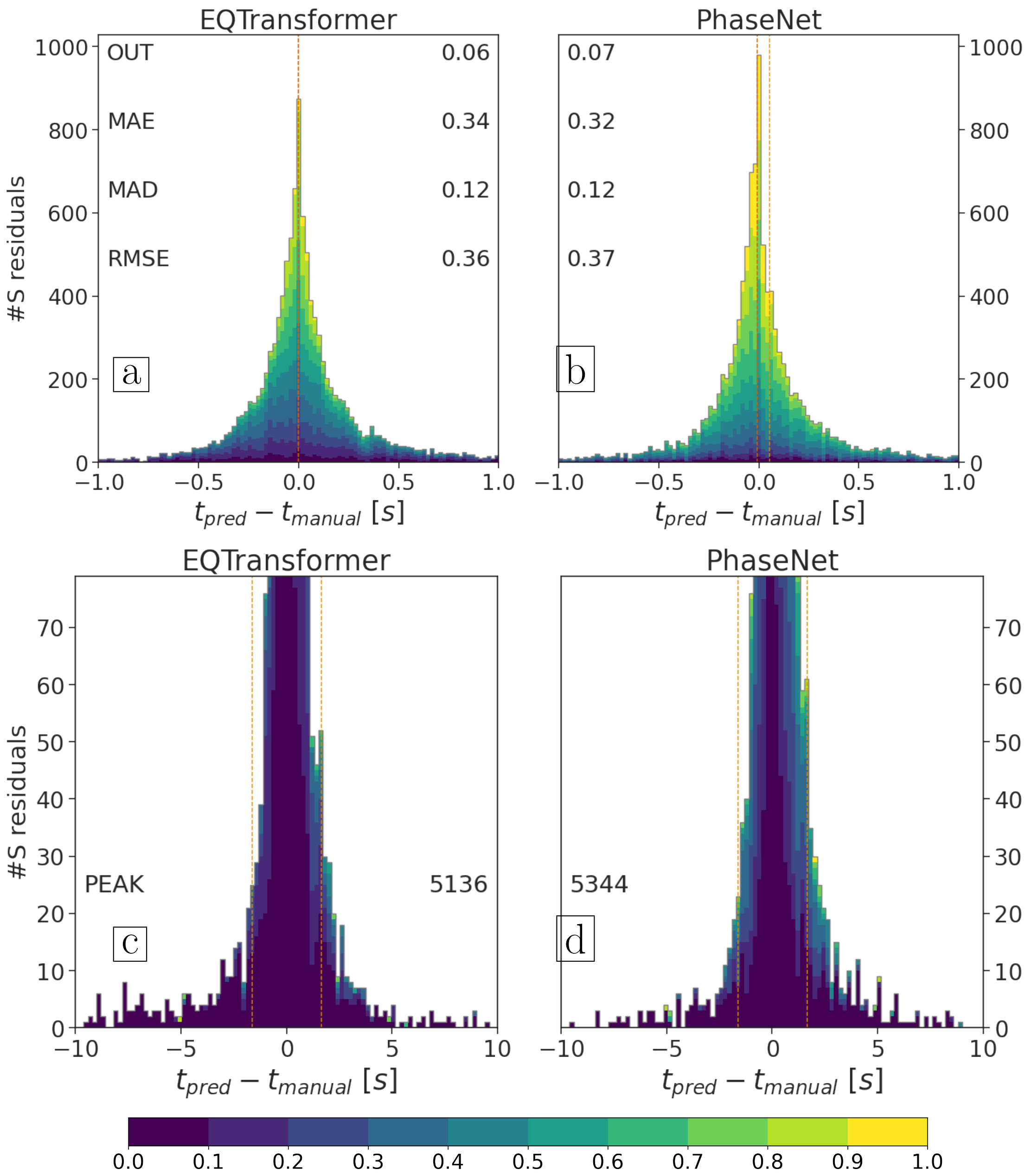


Figure S10: S residuals, pre-trained on STEAD.

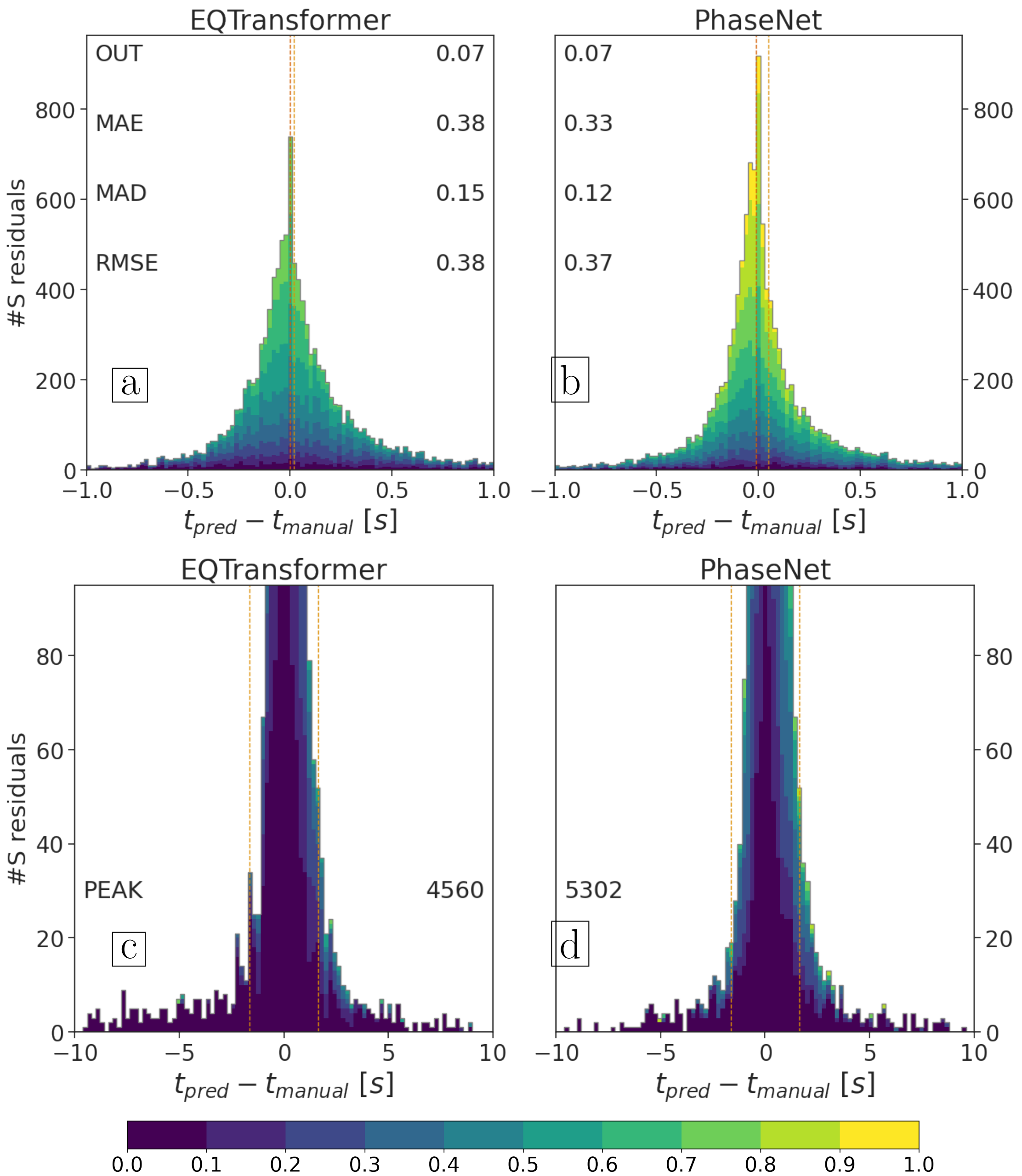


Figure S11: S residuals of models trained on OBS data without pre-training.

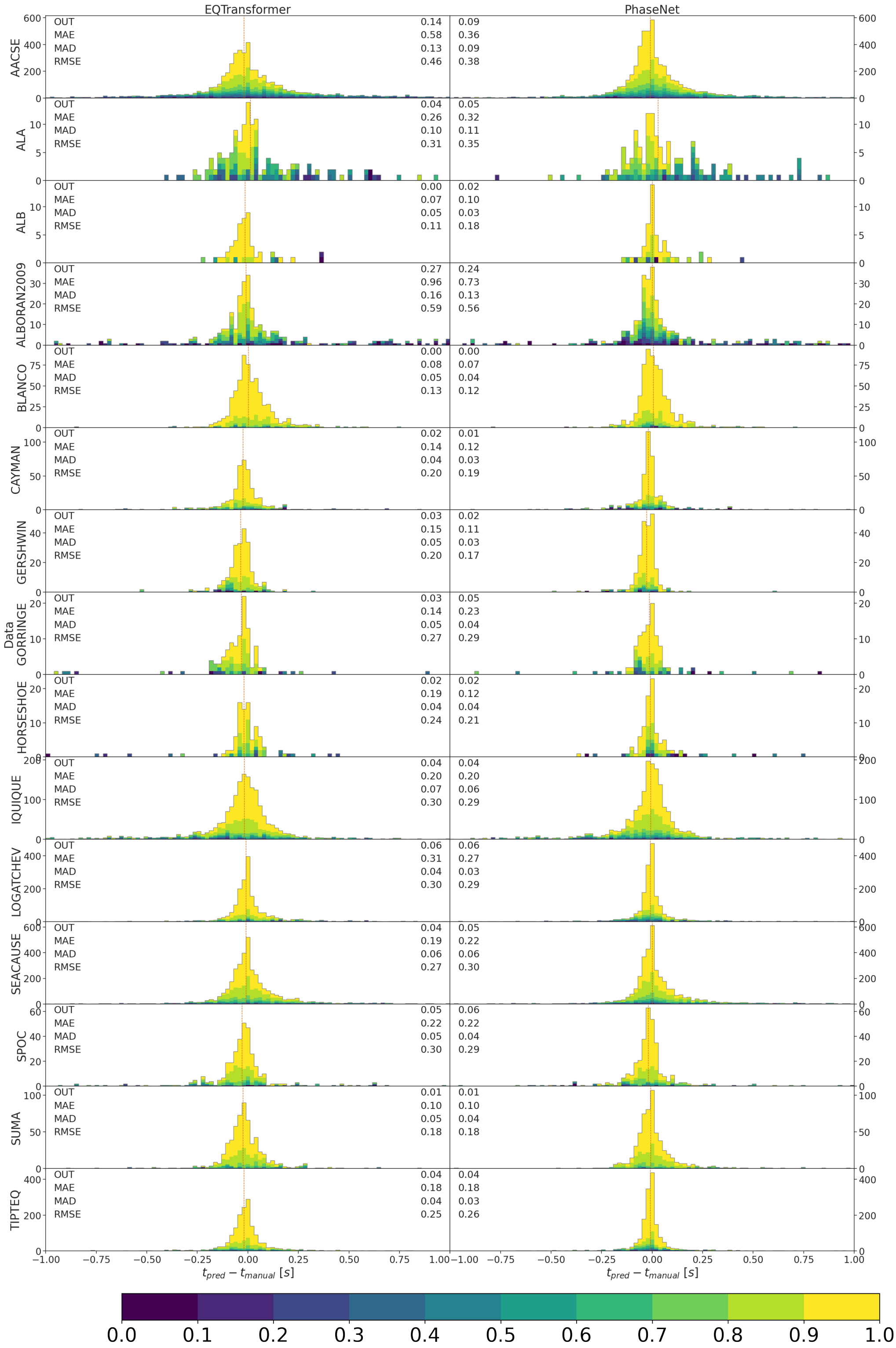


Figure S12: P residuals for individual experiments, pre-trained on STEAD

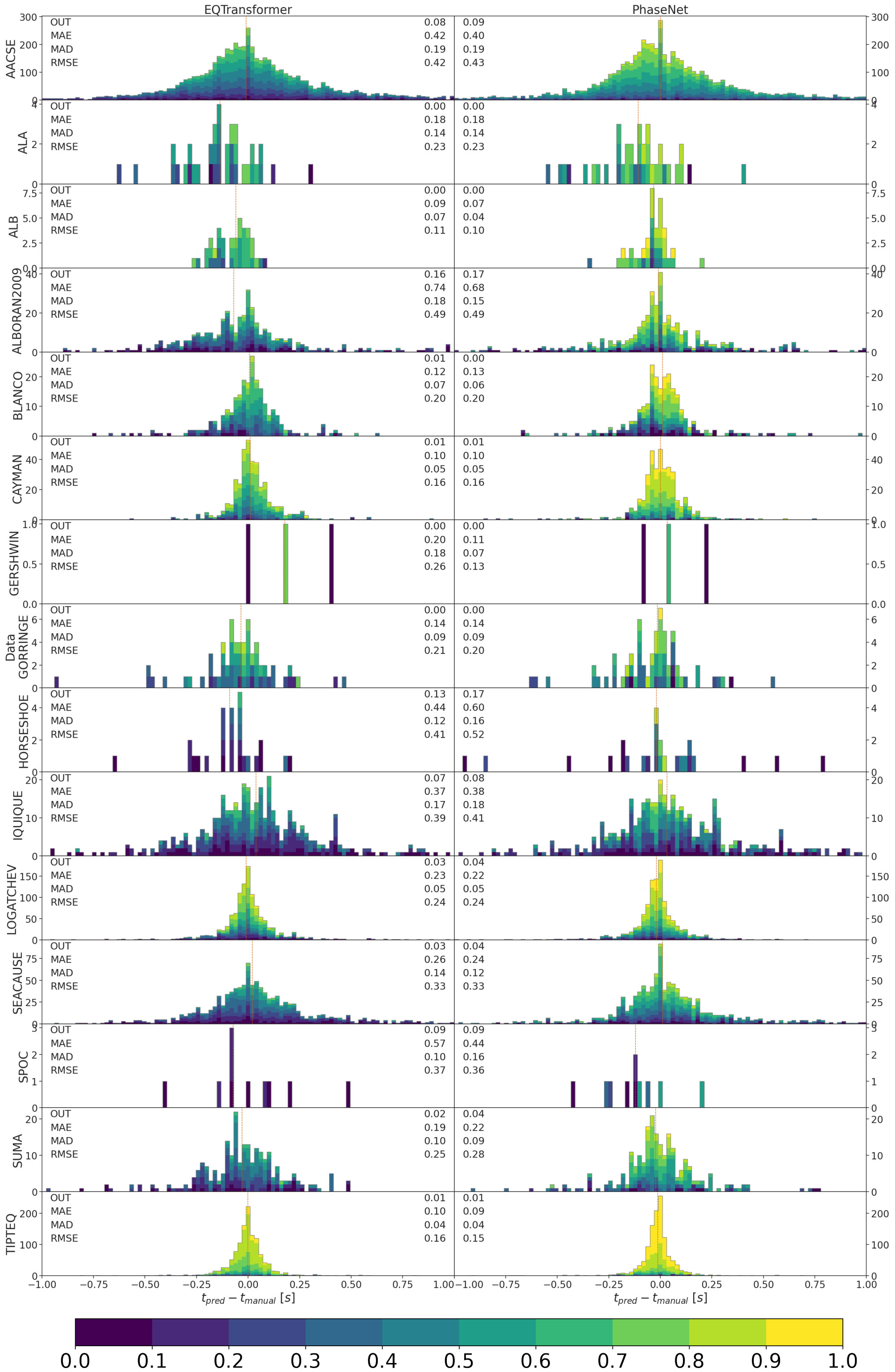


Figure S13: S residuals for individual experiments, pre-trained on STEAD

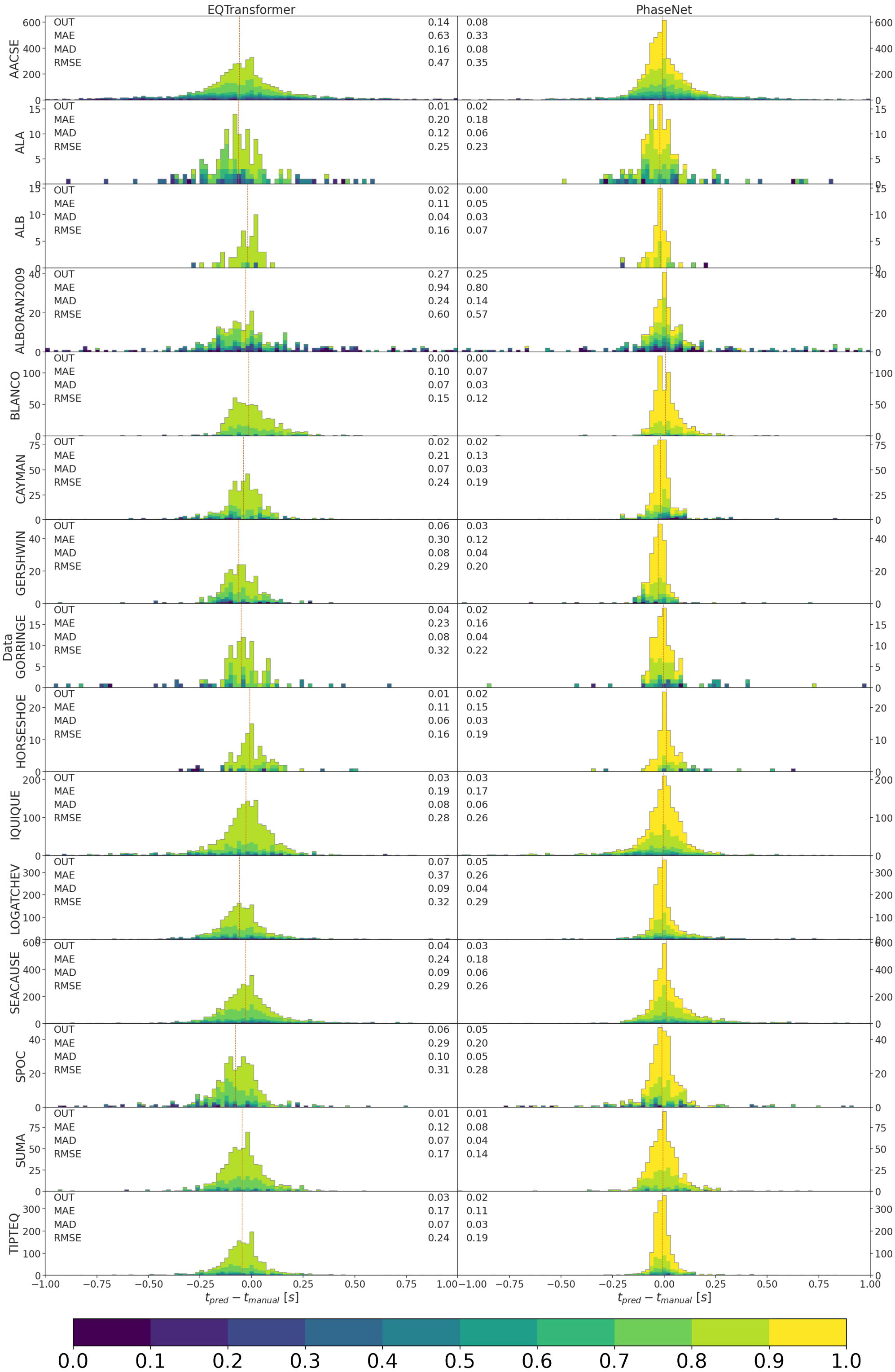


Figure S14: P residuals per experiment trained on OBS data without pre-training.

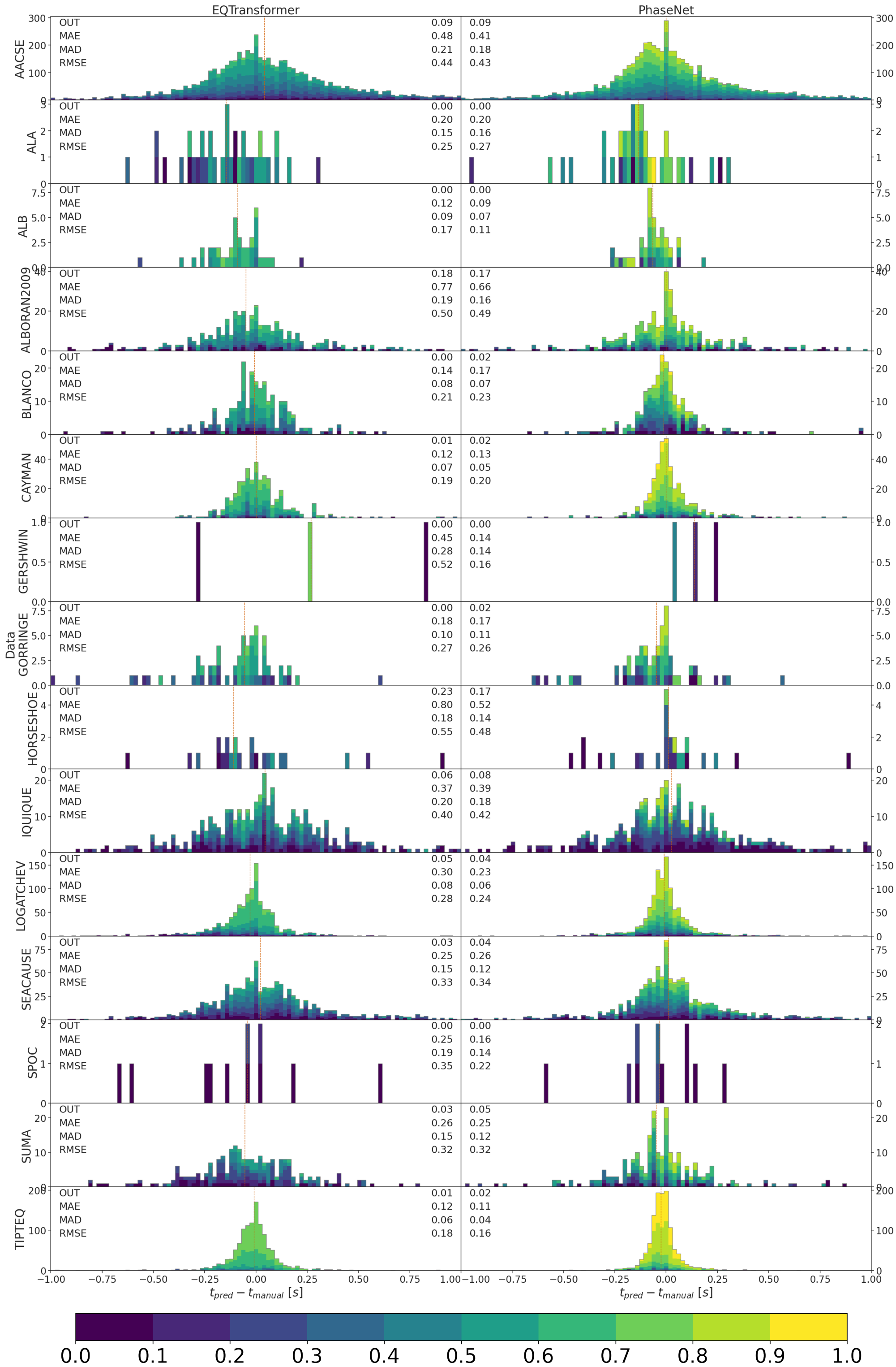


Figure S15: S residuals per individual experiment, trained on OBS data without pre-training.

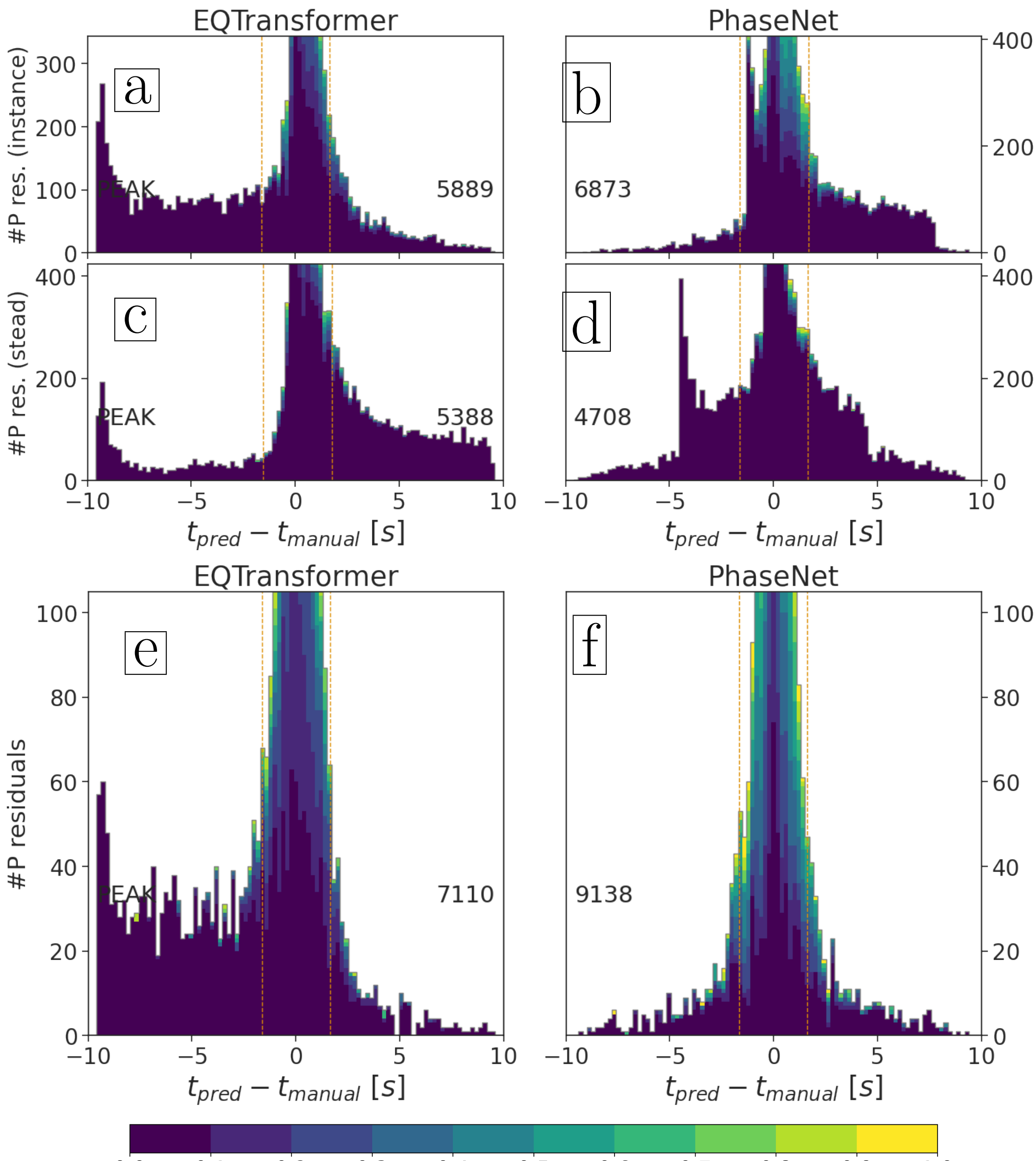
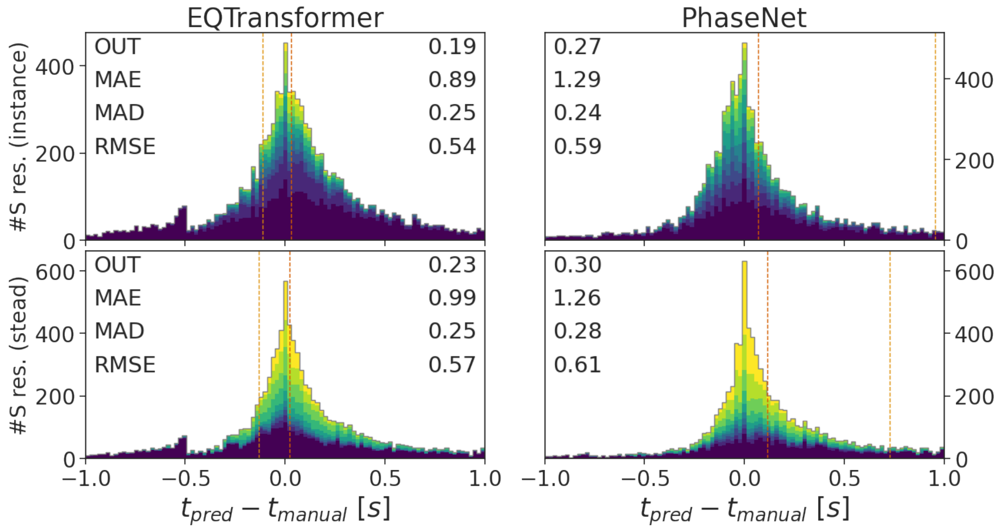


Figure S16: P residuals for three-component models. (a-d) trained only on land data using models trained on INSTANCE (a,b) and STEAD (c,d). (e-f) shows the result of training three-component models on the OBS dataset, but omitting the hydrophones, with pre-training on INSTANCE. As Fig. 10 but showing full range to ±10 s in order to highlight outliers.

(a)



(b)

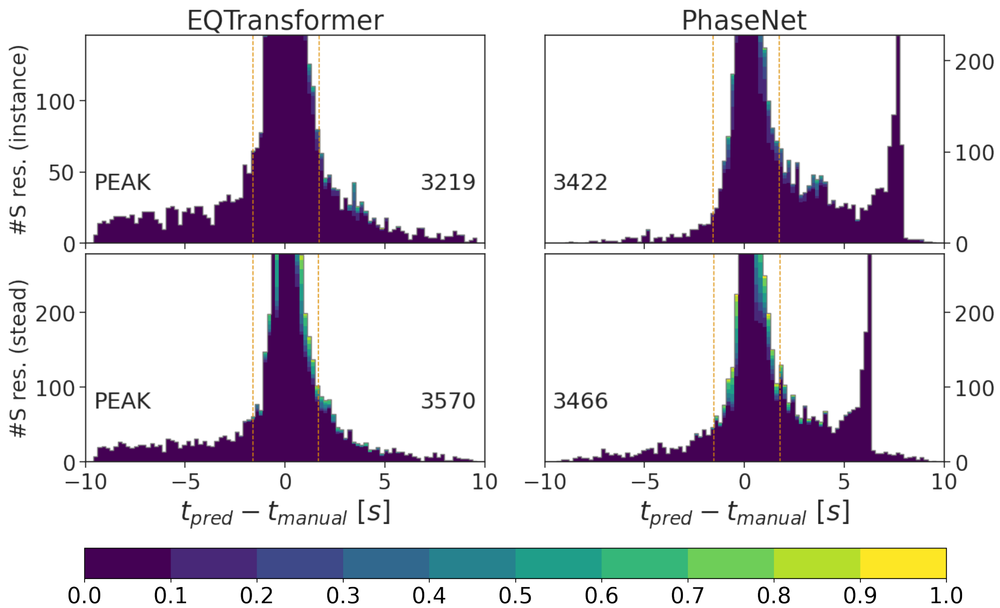


Figure S17: S residuals for three-component models, trained only on land data. (a) Range to ±1 s, (b) Range to ±10 s. Within each sub-figure: top–INSTANCE, bottom–STEAD). Hydrophone components were ignored.

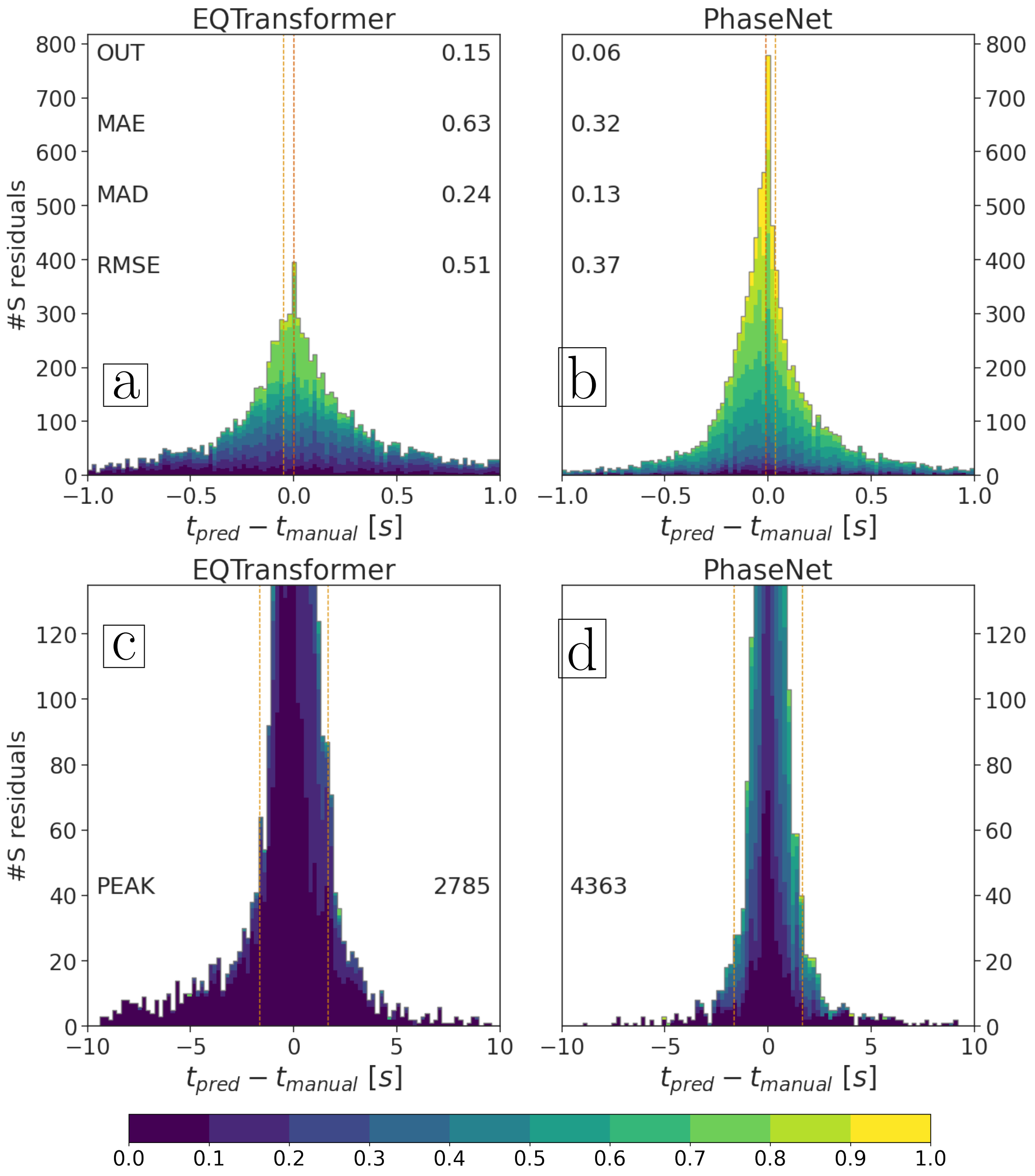


Figure S18: S residuals for three-component models, pre-trained on INSTANCE and then trained on OBS dataset. Figure is similar to Fig. 10a-b, but shows S residuals (instead of P residuals). Hydrophone components were ignored.

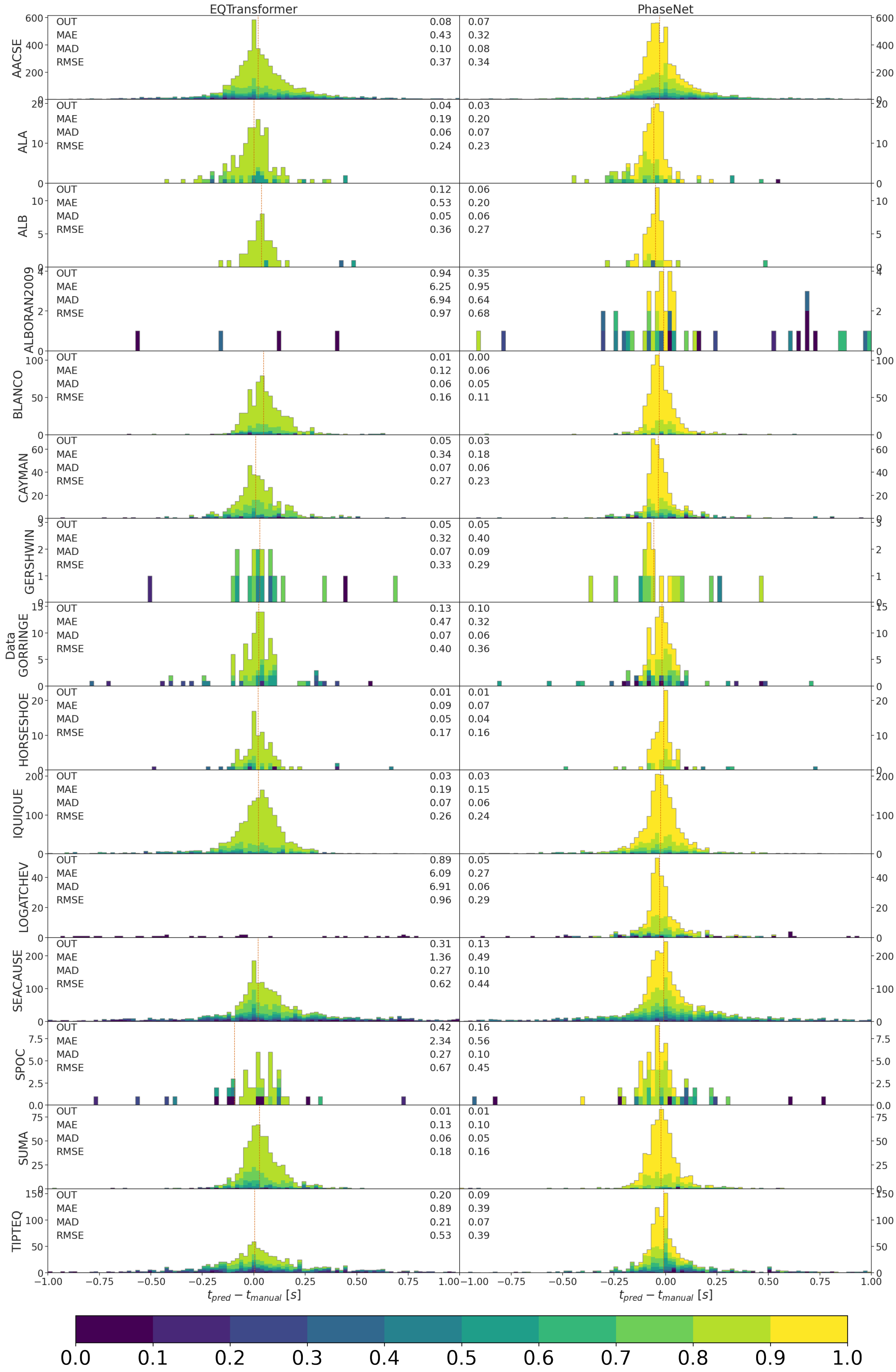


Figure S19: P residuals per experiment for three-component models, pre-trained on INSTANCE and then trained on OBS dataset. Hydrophone components were ignored.

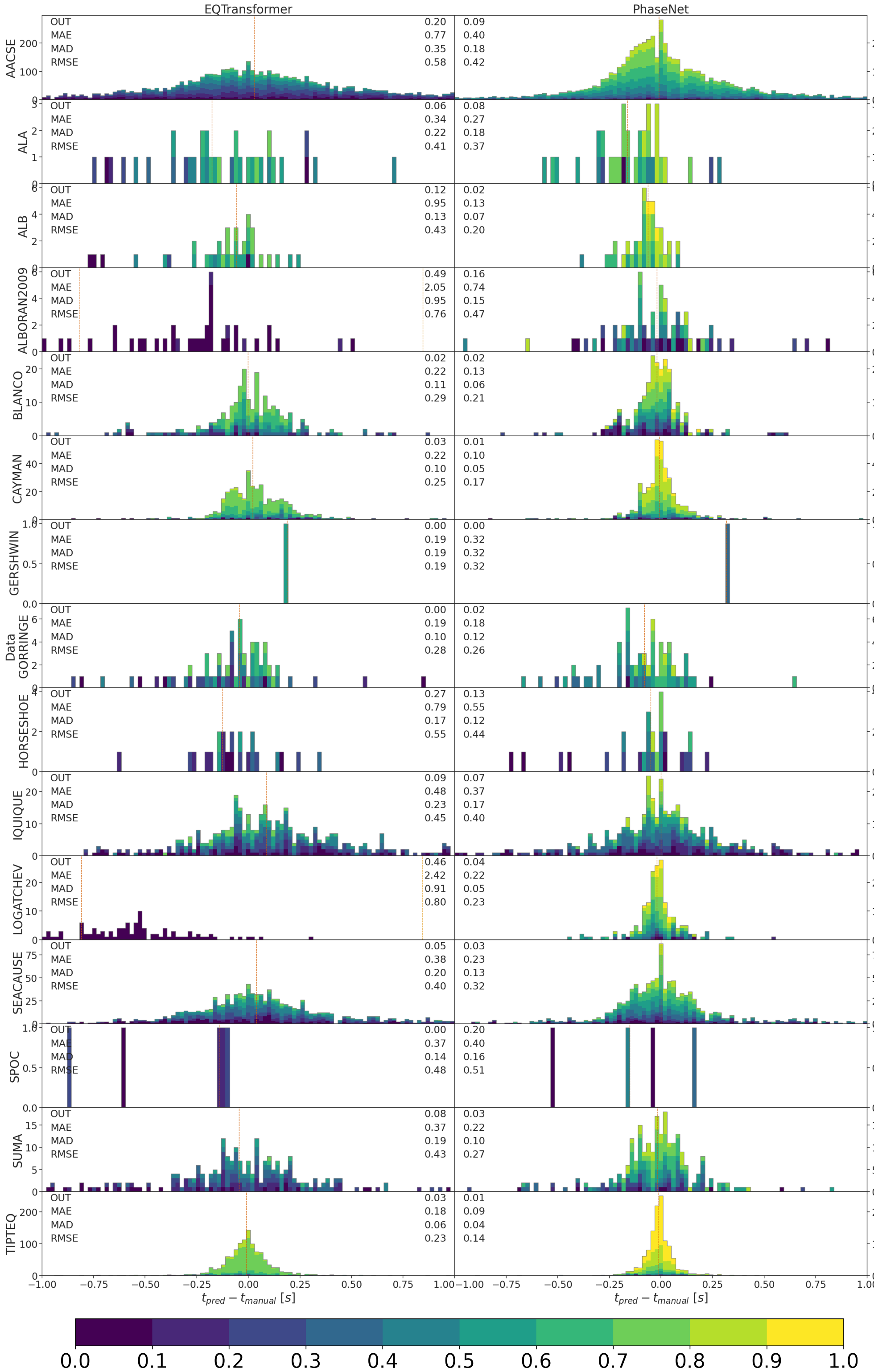


Figure S20: S residuals per experiment for three-component models, pre-trained on INSTANCE and then trained on OBS dataset. Hydrophone components were ignored.

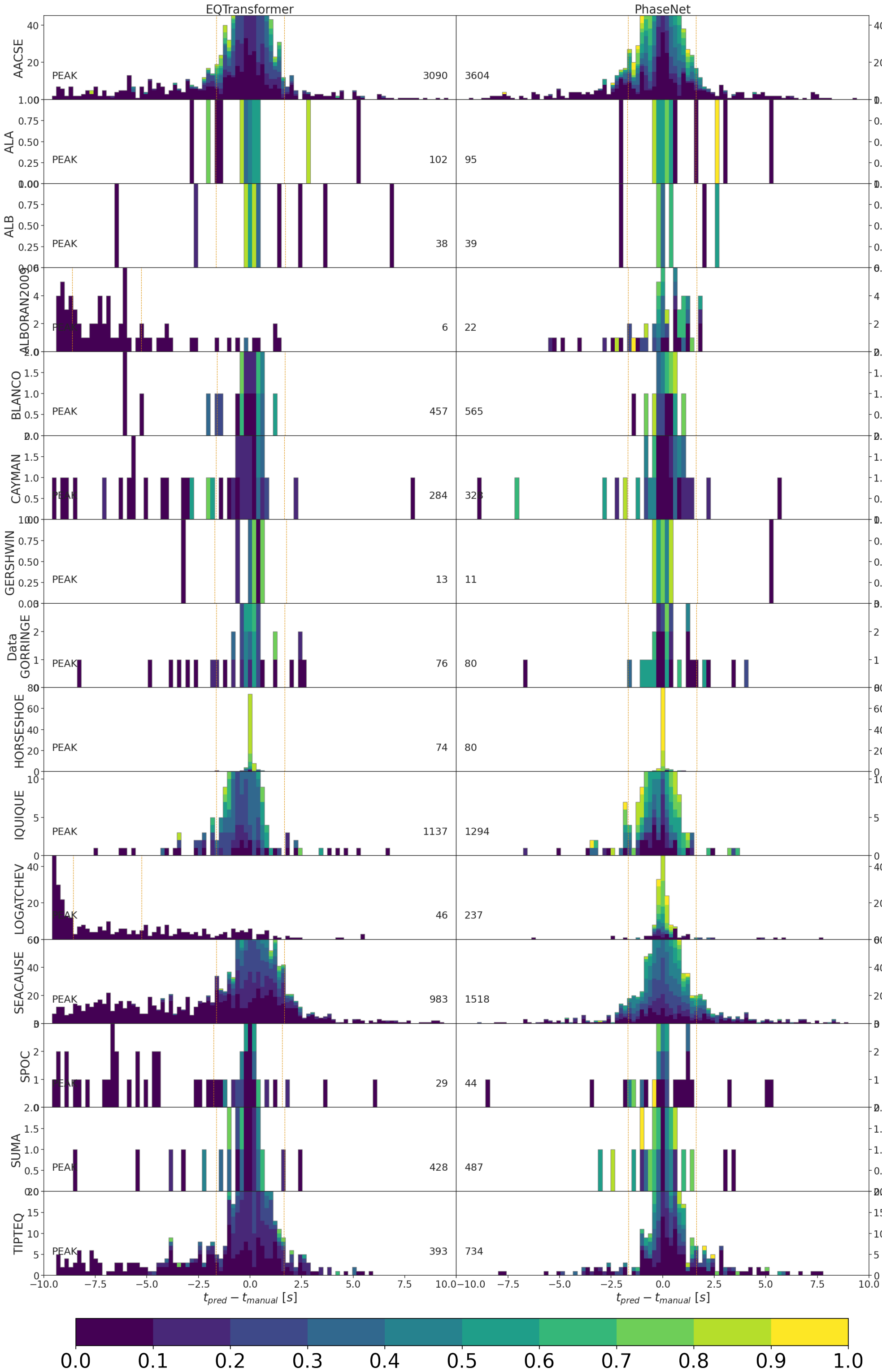


Figure S21: As Fig. S19 but showing full range to ±10 s in order to highlight outliers.

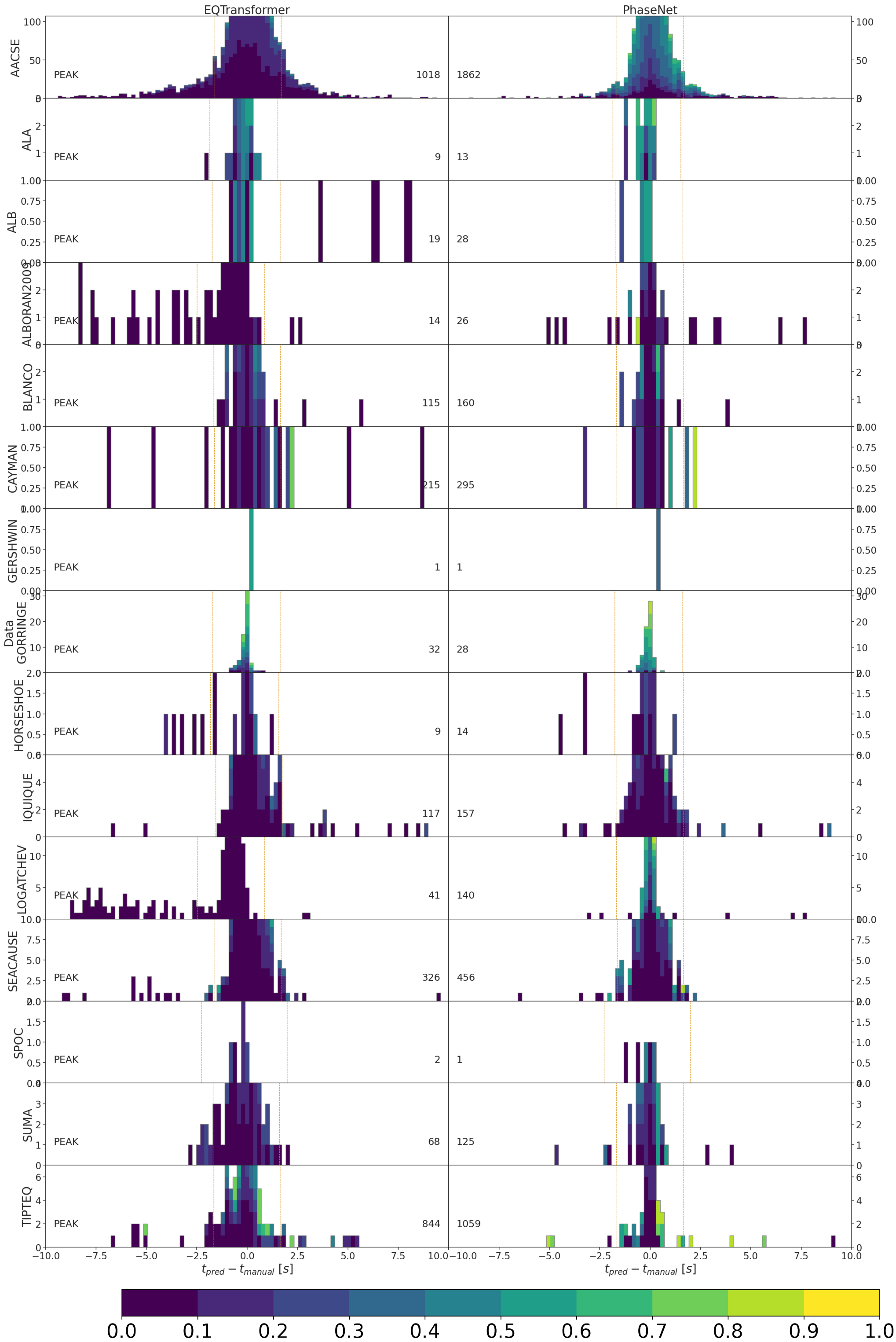


Figure S22: As Fig. S20 but showing full range to ±10 s in order to highlight outliers.