A. Appendix

This appendix contains a full account of our experimental results. These results correspond to the missing value mechanisms described in Section 4:

1. 10% MCAR (Figure 7), 30% MCAR (Figure 8) and 50% MCAR (Figure 9);
2. 30% MAR on 70% of the variables with a logistic masking model (Figure 10);
3. 30% MNAR generated with a logistic masking model, whose inputs are then themselves masked (Figure 11);
4. 30% MNAR on 30% of the variables, generated by censoring upper and lower quartiles (Figure 12).

These experiments follow the setup described in Section 4. In all the following figures, error bars correspond to ± 1 standard deviation across the 30 runs performed on each dataset. For some datasets, the W_2 score is not represented: this is due to their large size, which makes computing unregularized OT computationally intensive.

The results show that the proposed methods, Algorithm 1 and Algorithm 3 with linear and shallow MLP imputers, are very competitive compared to state-of-the-art methods, including those based on deep learning (Mattei & Frellsen, 2019; Yoon et al., 2018; Ivanov et al., 2019), in a wide range of missing data regimes.

Runtimes. Figure 6 represents the average runtimes of the methods evaluated in Figure 11. These runtimes show that Algorithm 1 has computational running times on par with VAEAC, and faster than the two remaining DL-based methods (GAIN and MIWAE). Round-robin methods are the slowest overall, but the base imputer model being used seems to have nearly no impact on runtimes. This is due to the fact that the computational bottleneck of the proposed methods is the number of Sinkhorn batch divergences that are computed. This number can be made lower by e.g. reducing the number of gradient steps performed for each variable (parameter K in algorithm 3), or the number of cycles t_{max}. This fact suggests that more complex models could be used in round-robin imputation without much additional computational cost.

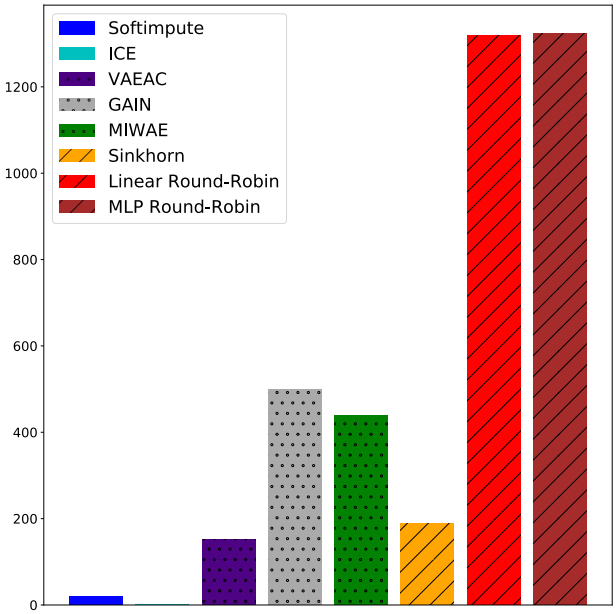

Figure 6: Average runtimes (in seconds, over 30 runs and 23 datasets) for the experiment described in fig. 11. Note that these times are indicative, as runs where randomly assigned to different GPU models, which may have an impact on runtimes.

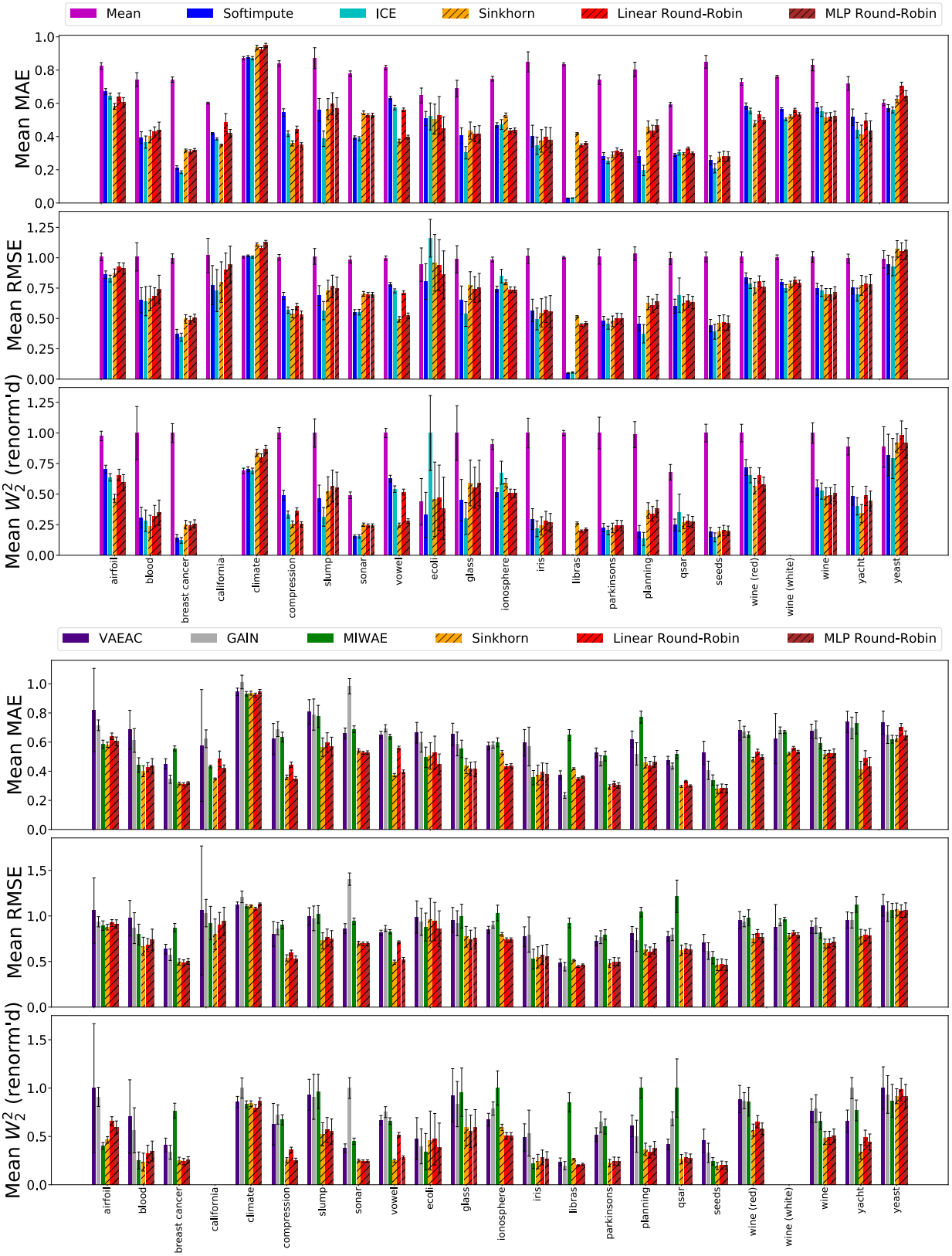

Figure 7: (**10 % MCAR**)

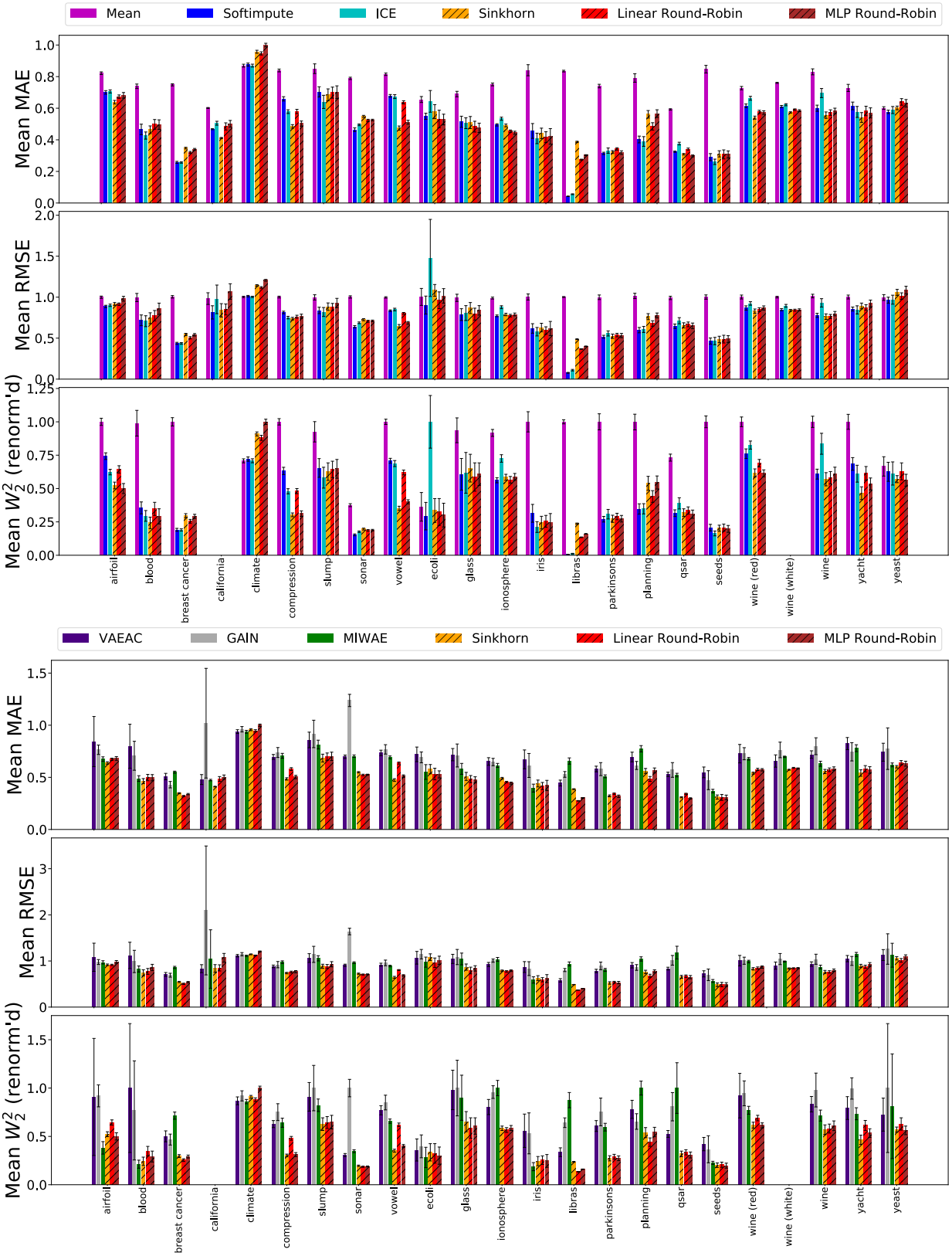

Figure 8: (**30 % MCAR**)

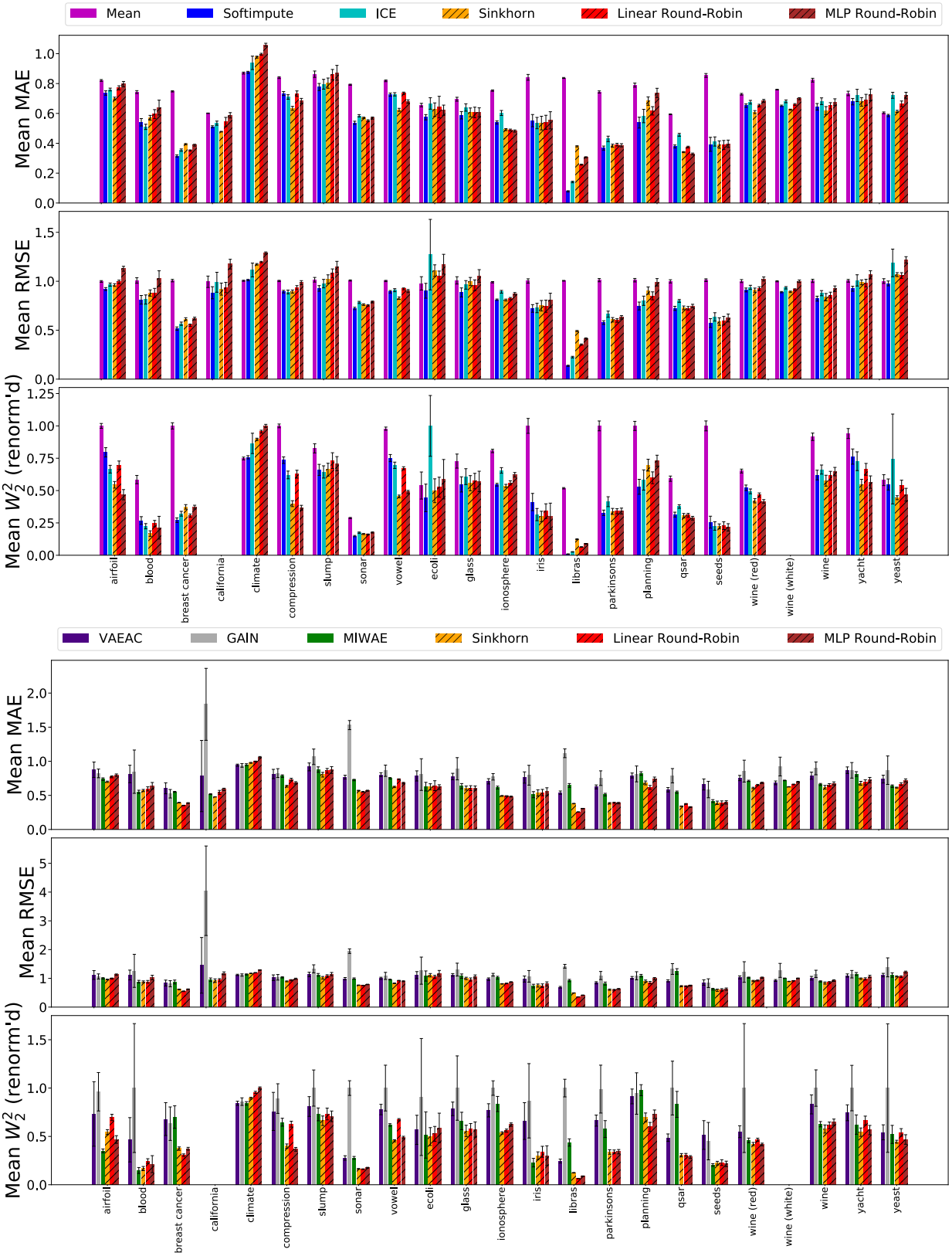

Figure 9: (**50 % MCAR**)

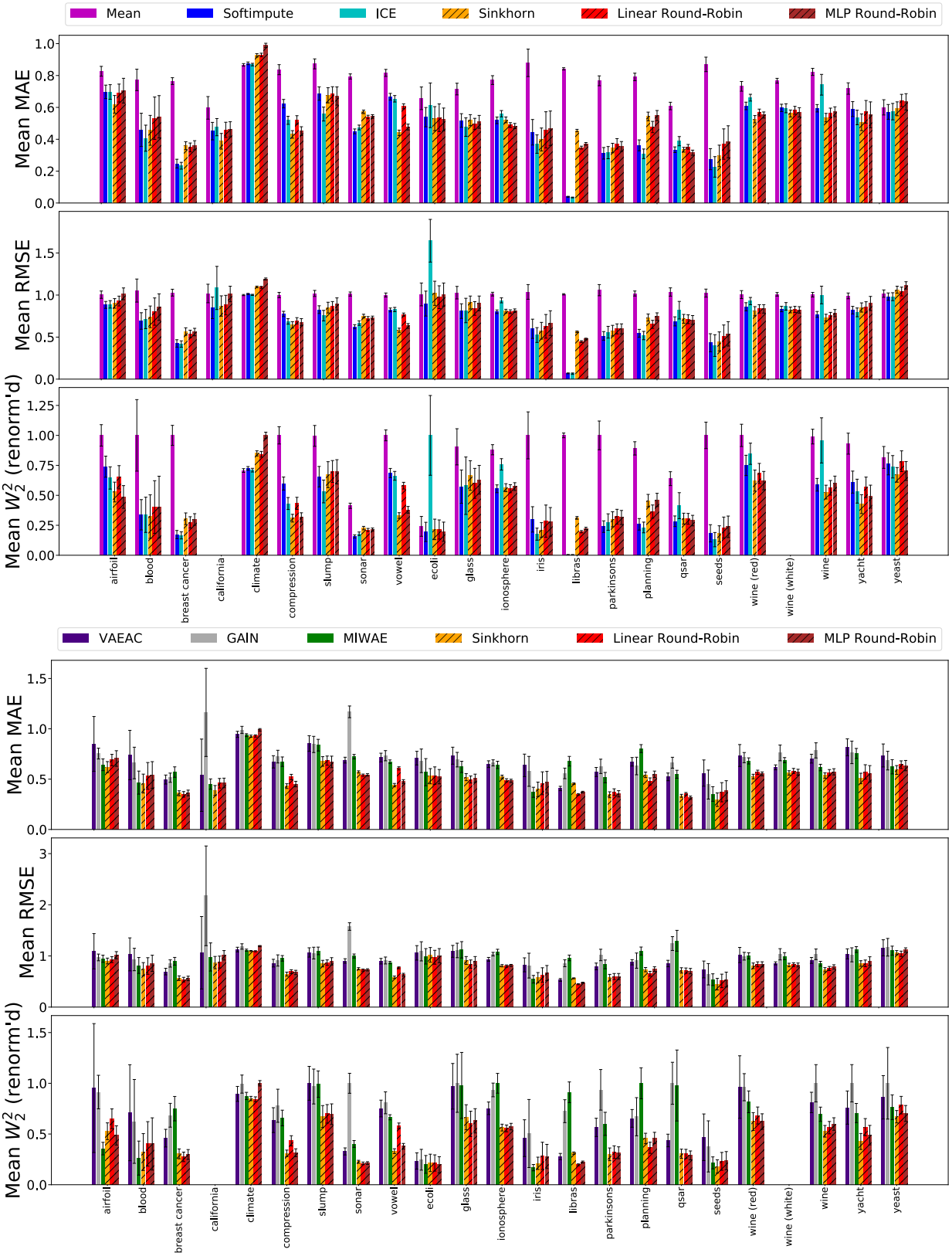

Figure 10: (**30 % MAR**)

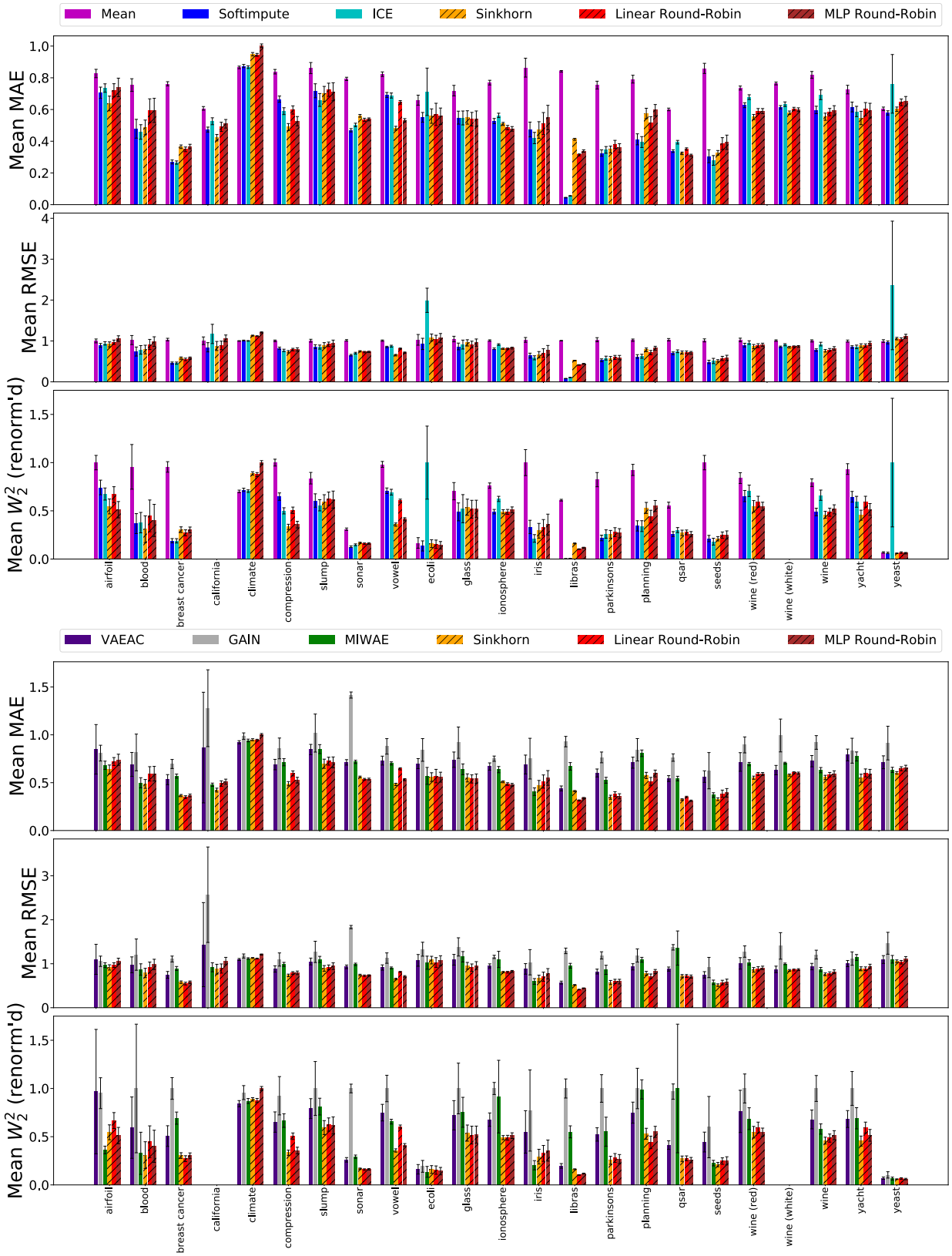

Figure 11: (**30 % MNAR, logistic masking**)

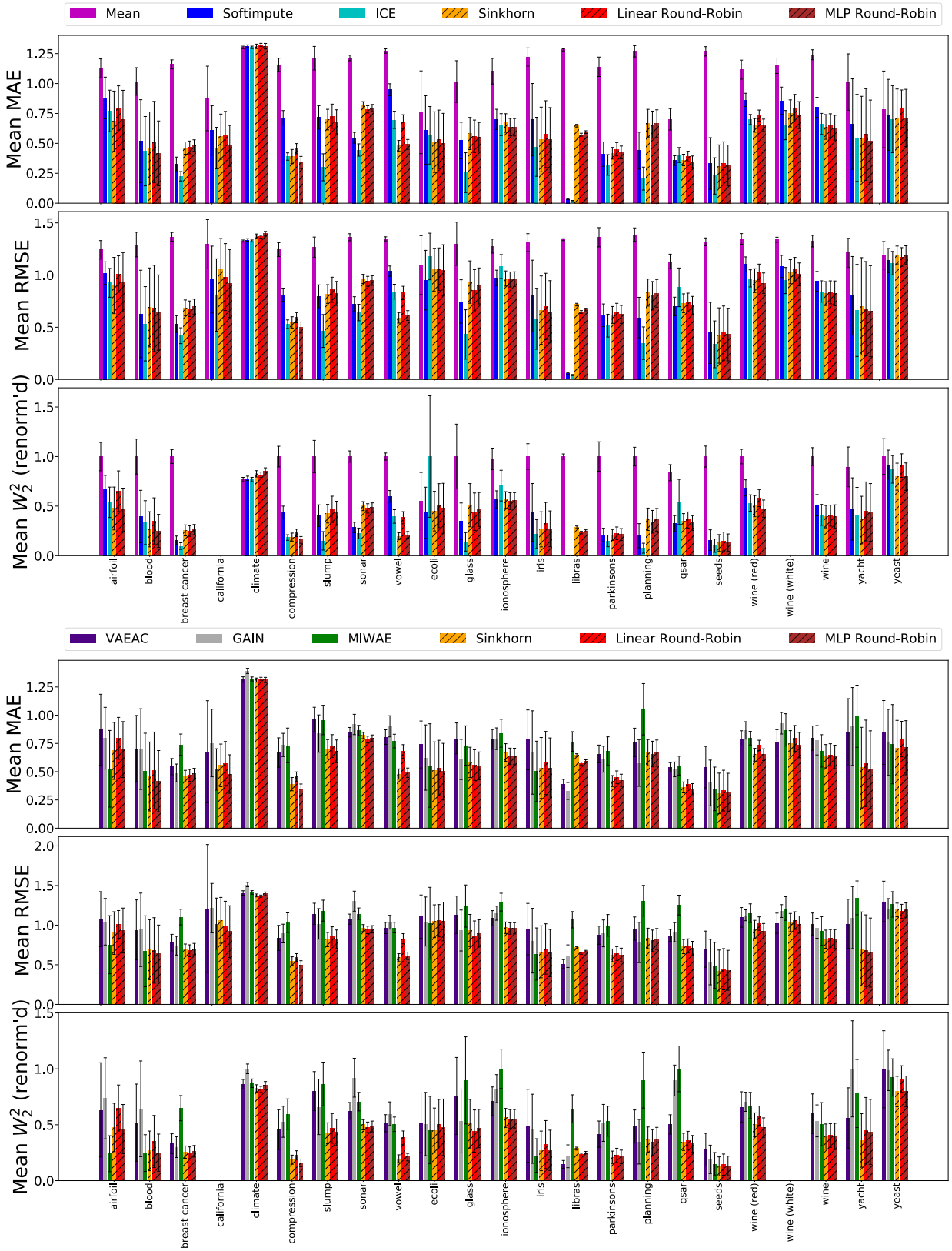

Figure 12: (**30 % MNAR, quantile masking**)