

Data leakage jeopardizes ecological applications of machine learning

In the format provided by the
authors and unedited

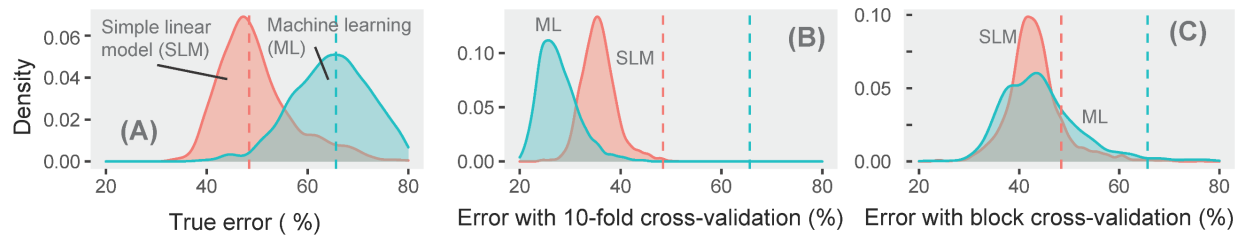


Fig. S1. Distribution of actual and estimated errors from simple linear and basic machine learning models for 2,000 artificial data sets (new visualization of data from the third example study in the main text; Stock, 2022, *ISPRS Journal of Photogrammetry and Remote Sensing*, <https://doi.org/10.1016/j.isprsjprs.2022.02.023>). The artificial data mimic typical data for supervised learning applications in satellite remote sensing of marine biological phenomena. Each plot can be interpreted like a smooth histogram. The x-axis shows prediction errors (expressed as mean absolute differences in % between predictions and observations). Vertical lines are the medians of the models' true errors. **(A)** Simple linear models had the lowest prediction errors (as absolute percentage difference) on most artificial data sets. **(B)** In contrast, errors estimated with 10-fold cross-validation were smaller, suggesting that a machine learning model (random forests) had the lowest errors. **(C)** Spatial block cross-validation reduced but did not eliminate this underestimation of prediction errors, especially for machine learning models.