Benchmarking and Comparative Evaluation
Volume 1 · Number 1 · 2026

The Same Wind Forecasts, Different Winners under Absolute and Relative Loss

Spencer Palmer[1] · Charles Hughes[2] · Paul Lewis[3]

[1] Poland · [2] Hungary · [3] Greece

Original publication: October 6, 2026 | Article 651

Abstract

Persistence ranked behind a neighboring-input regression under absolute wind-speed error but ahead under target-scaled and symmetric relative losses. All comparisons used the same 51,384 forecast origins and models fitted before the held year. Across 361 dates, neighbor-minus-persistence MAE was -0.011618 m/s. Under a target-denominator floor of 0.1 m/s, the contrast became 0.023236 in dimensionless loss; a 0.5 floor reduced it to 0.001753 without restoring the absolute-error ranking. The source contained 692 reported zero targets, making raw percentage error undefined for the full sample. This is a change in the evaluation criterion, rather than changed predictions or evidence that one metric supplies a universally correct forecasting objective.

An evaluation-criterion audit using frozen predictions and a common chronological holdout from one public wind farm.

The Same Wind Forecasts, Different Winners under Absolute and Relative Loss

An evaluation-criterion audit using frozen predictions and a common chronological holdout from one public wind farm.

Abstract

Persistence ranked behind a neighboring-input regression under absolute wind-speed error but ahead under target-scaled and symmetric relative losses. All comparisons used the same 51,384 forecast origins and models fitted before the held year. Across 361 dates, neighbor-minus-persistence MAE was -0.011618 m/s. Under a target-denominator floor of 0.1 m/s, the contrast became 0.023236 in dimensionless loss; a 0.5 floor reduced it to 0.001753 without restoring the absolute-error ranking. The source contained 692 reported zero targets, making raw percentage error undefined for the full sample. This is a change in the evaluation criterion, rather than changed predictions or evidence that one metric supplies a universally correct forecasting objective.

1. Benchmark rankings depend on the quantity being minimized

A forecasting table usually names a winning method after averaging an error score. The method can appear to change quality when the averaging rule changes, even though every prediction and target remains identical. Absolute error measures a discrepancy in the original physical unit. Relative scores additionally depend on the scale of the reported target, or on both target and prediction. Those choices can prioritize different parts of a wind-speed record and change which model receives the smallest average.

We ask whether three fixed wind forecasts keep the same ranking under explicitly defined absolute, root-square, target-scaled, and symmetric losses. The primary paired comparison is a neighboring-input regression minus persistence. The proposition of rank invariance predicts the same sign across the selected criteria. Its alternative is a criterion-dependent sign change on the same observations. This question concerns measurement and evaluation, without claiming that a preferred metric can be inferred from which model it rewards.

The benchmark-comparison problem is relevant to wind prediction such as Yan and Kuruoglu, 2023. Our forecasts do not implement the cited graph-sign method, and no new robustness or speed claim is made for it. We use simple frozen regressions and persistence to make the evaluation object inspectable. A change in their reported ordering can inform how a later benchmark is specified; it cannot determine the ordering of algorithms that have not been executed.

2. Three predictions on one shared later-year sample

The data are observed La Haute Borne SCADA wind-speed readings captured in ENGIE / NREL OpenOA contributors, 2026 and distributed with the example resource associated with Perr-Sauer et al., 2021. The original table has 420,480 rows across four turbines in 2014 and 2015. We discard ninety-six rows belonging to conflicting duplicate turbine-time keys before forming a wide time series. Every evaluation origin requires current readings from all four turbines and an observed R80711 target exactly ten minutes later.

The own-only model has an intercept and the target turbine's current speed. The neighboring-input model adds current readings from the other three turbines. Both are ordinary least-squares fits on the 52,321 origins whose targets occur before 2015. Persistence repeats the current own-turbine speed without an

estimated coefficient. These rules are frozen before the final year, with no metric-specific refitting. There are 51,384 common origins in the 2015 holdout, so differences among criteria do not come from changing the primary sample.

Operational forecasting considerations in Pinson, 2013 and the wind-methods review in Foley et al., 2012 motivate specifying the horizon and available information. The spatial-resource context in Haslett and Raftery, 1989 supports neighboring observations as plausible predictors, not a causal interpretation of their coefficients. We predict wind speed, not electrical power. An advantage in meters per second or a percentage-style loss does not establish a change in energy output, operating cost, or revenue without additional information.

3. Reported zeros require a visible decision

Among held targets, 692 are reported as exactly zero, 787 are below 0.1 m/s, 1,193 below 0.5, and 1,717 below one. We retain these values as observations from the source rather than automatically converting them into missing data. Their physical provenance is not independently verified in this analysis. A reported zero can affect a loss formula regardless of whether it reflects genuine conditions, instrument behavior, or another source-specific recording process.

Ordinary absolute percentage error divides the absolute prediction discrepancy by the target magnitude. At a zero target that expression is undefined, including a zero-over-zero case for a perfectly matching zero forecast. Consequently a finite full-sample mean requires an explicit convention or a changed sample. We do not silently delete those targets and present the result as the same complete holdout. A positive-target-only descriptive calculation is preserved separately, with its smaller denominator and different origin weighting clearly identified.

The accuracy-measure discussion in Hyndman and Koehler, 2006 and point-forecast evaluation framework in Gneiting, 2011 help explain why this is an objective-definition problem. Replacing an undefined denominator is not a neutral software repair that leaves the original criterion unchanged. It creates a specified alternative loss, with different relative priorities across observations. Whether that alternative suits a real application requires a substantive purpose for the score rather than a preference for one method's resulting rank.

4. Five explicit summaries

The first score is ordinary absolute wind-speed error. Within each date we average discrepancies, then give every retained date equal weight. The second computes each date's root mean squared error and subsequently averages those daily roots. This is not the root of pooled squared error across all origins, nor the root of the average daily mean squared errors. The placement of the square root is part of the estimand and is preserved in the output records.

The comparison between absolute and squared-error summaries discussed by Willmott and Matsuura, 2005, Chai and Draxler, 2014, and Hodson, 2022 motivates keeping their units and interpretations distinct. We do not claim that these sources establish one universally preferred loss. Both summaries remain in meters per second, while emphasizing discrepancies differently. Their agreement in ranking is an empirical observation for these forecasts, not a mathematical requirement for every forecast distribution or another aggregation scheme.

Two target-scaled scores divide each origin's absolute error by the greater of its reported target speed and a fixed floor. The floors are 0.1 and 0.5 m/s. Each resulting ratio is averaged within a date and then across dates. These are explicitly floored ratios, not raw MAPE. The floors are sensitivity conventions chosen for

this analysis; they are not asserted to be instrument detection limits, industry standards, or validated tolerances for wind operations.

The fifth score uses twice absolute error divided by the sum of absolute prediction and absolute target. When both are zero its contribution is defined as zero. The symmetric-MAPE discussion in Goodwin and Lawton, 1999 motivates stating this formula rather than relying on its name to imply every desirable symmetry. Unlike the target-floored ratios, its denominator depends on the prediction as well as the target. It cannot be interpreted simply as a shared observation-weighted absolute error across models.

5. Date weighting and the limits of a benchmark label

After within-date calculation there are 361 available dates. Partial dates remain, and each receives equal weight. That choice avoids letting more densely observed dates automatically dominate, but it does not estimate error for the four dates with no eligible origins. It also leaves availability selection unresolved. All five primary summaries use the same dates and source origins; we change the loss definition, not the information available to a forecast or the inclusion rule for its targets.

The competition perspective in Makridakis and Hibon, 2000 and out-of-sample evaluation discussion in Tashman, 2000 motivate presenting an evaluation protocol with a ranking. A benchmark name alone does not identify the loss, temporal boundary, or aggregation unit. Temporal-validation concerns in Bergmeir et al., 2018 and Roberts et al., 2017 similarly explain why this later-year comparison answers a different question from a random split of neighboring origins. We do not retune a model after seeing which score favors it.

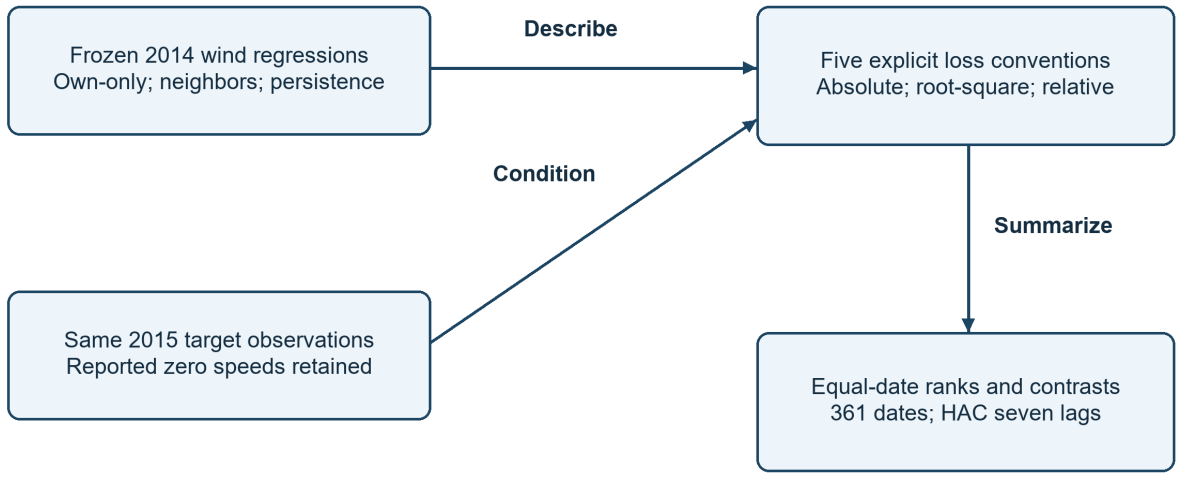

Figure 1. Identical later-year predictions are scored using five declared criteria; paired daily contrasts compare the neighboring model with persistence.

Predictive scoring principles in Gneiting and Raftery, 2007 reinforce that a numerical loss defines what an evaluation rewards. Calibration questions in Guo et al., 2017 concern another property and are not answered by a small deterministic error. The figure therefore contains no calibrated uncertainty distribution or claim about probabilistic wind forecasting. Its outputs are observed point losses and their dated comparisons, with the scale of each score stated explicitly.

6. Point rankings reverse without any prediction changing

Table 1 gives the equal-date means. Neighboring-input regression has the smallest absolute and daily-root-square error. Persistence has the smallest value under both target-floored ratios and the stated symmetric score. The own-only regression does not win any of these point comparisons. This is a rank reversal on a common dataset, not a comparison of scores measured on different targets or a result obtained by modifying a forecast after inspecting the criterion.

Table 1. Equal-date scores for the same held-year forecasts; smaller is better within a row.

Criterion	Own-only	Neighboring inputs	Persistence	Unit
Absolute error	0.433715	0.423470	0.435088	m/s
Average daily RMSE	0.595033	0.576177	0.599246	m/s
Target-floor ratio, 0.1	0.172957	0.170062	0.146826	Dimensionless
Target-floor ratio, 0.5	0.117872	0.114607	0.112854	Dimensionless
Stated symmetric score	0.128697	0.126071	0.116930	Dimensionless

The relative-score rows are fractions. Multiplying them by one hundred would express a percentage-style scale but would not turn them into meters per second or make their differences comparable with absolute error. We do not average these different score types into a composite ranking. Such a composite would need its own weighting rationale and could obscure precisely the evaluation distinction this experiment is designed to expose.

7. Paired mean differences have opposite signs

For absolute error, each target-floor ratio, and the symmetric score, we subtract persistence loss from neighboring-model loss within each date. Separate intercept-only ordinary least-squares fits estimate these mean contrasts. Each uses 361 rows and HAC covariance with seven retained-date lags, a finite-sample correction, and a t reference with 360 degrees of freedom, following Newey and West, 1987. The framework in Diebold and Mariano, 1995 and West, 1996 supports pairing predictive losses on common evaluation units.

Table 2. Neighbor-minus-persistence daily loss contrasts; the first model is in m/s and the others are dimensionless.

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	-0.011618 (0.001847)	[-0.015251, -0.007986]	361
M1: fit	R-squared = 0.0000	-	-	361
M2	Mean outcome/contrast	0.023236 (0.003767)	[0.015827, 0.030645]	361
M2: fit	R-squared = 0.0000	-	-	361
M3	Mean outcome/contrast	0.001753 (0.000719)	[0.000339, 0.003167]	361
M3: fit	R-squared = 0.0000	-	-	361
M4	Mean outcome/contrast	0.009141 (0.002637)	[0.003955, 0.014327]	361
M4: fit	R-squared = 0.0000	-	-	361

Model labels: M1 = MAE: neighbor minus persistence; M2 = Target-scaled loss, 0.1 m/s floor: neighbor minus persistence; M3 = Target-scaled loss, 0.5 m/s floor: neighbor minus persistence; M4 = SMAPE: neighbor minus persistence.

The absolute contrast is -0.011618084 m/s, with standard error 0.001847157 and interval -0.015250659 to -0.007985510. Under the 0.1 target floor it becomes 0.023236065, standard error 0.003767271 and interval 0.015827442 to 0.030644688. Both the sign and the loss scale change. Their uncertainty intervals exclude zero under the declared approximation, supporting a criterion-dependent ordering for this resource rather than a difference arising only from small point-score rounding.

For the 0.5 floor, the contrast is 0.001753317 with interval 0.000339408 to 0.003167226. The symmetric-score contrast is 0.009141202, with interval 0.003955071 to 0.014327333. Increasing the floor greatly reduces the relative-score disadvantage but does not reverse its sign here. The daily-root-square row and the full three-way rankings remain descriptive; we do not attach an uncomputed interval to every pairwise entry in Table 1.

8. The score changes the priority given to reported low speeds

For the target-floored criteria, the denominator determines an inverse weighting of absolute error. An equal absolute discrepancy receives more influence when the reported target is small, until the selected floor limits that influence. This is a mathematical feature of the score. The extent to which it changes a particular model ranking is empirical and depends on the joint distribution of predictions and reported speeds. We do not assume that every relative-error benchmark must favor persistence.

The 692 zero targets make the convention especially visible, while nonzero small targets also contribute to the reweighting. We have not separately decomposed the ranking reversal into zero, low-positive, and other observations, so we do not attribute the entire reversal to zeros alone. The symmetric score adds a prediction-dependent denominator, creating another criterion change. Similar point rankings under these relative criteria do not make their numerical values or observation-level priorities identical.

Measurement-error considerations in Carroll et al., 2006 are relevant to the source's reported speeds, but the calculation cannot certify their physical accuracy. We neither relabel zero readings as verified calm weather nor dismiss them as recording faults. A quality-control investigation could change the intended evaluation population, requiring a new documented analysis. The current comparison retains the values and makes their mathematical effect on undefined or floored ratios explicit.

9. A ranking is conditional on purpose and dependence assumptions

Heteroskedasticity correction in White, 1980 alone would not represent serial dependence among dates, while the dependence discussion in Cameron and Miller, 2015 also limits interpreting one farm as multiple independent sites. Seven available-date lags remain a covariance choice, especially when dates are absent. The intervals do not cover other farms, horizons, retraining samples, or possible source quality-control revisions. Their R-squared values are zero because each model estimates a mean, not a forecast relationship.

Model-selection bias in Varma and Simon, 2006 would arise if a criterion were chosen because it made a favored forecast win and then described as independently validated. Influence concerns in Cook, 1977 motivate preserving unusual dates, rather than removing them to recover a preferred ordering. Multiplicity considerations in Benjamini and Hochberg, 1995 also constrain broad statements from several pointwise comparisons. We report the chosen criteria and observed signs, without claiming that their collection determines an optimal decision rule.

Interpretation guided by Wasserstein and Lazar, 2016 requires both physical magnitude and criterion meaning. The neighboring model's absolute-error advantage is small in m/s, while persistence's relative-score advantage answers a differently weighted question. Neither result identifies operational cost

or a universal definition of robustness. An application would need to specify which discrepancies matter and how reported low speeds should enter that objective before the benchmark ranking could justify a practical model choice.

10. Preserve the loss definition with the forecast table

The completed audit finds a clear reversal between absolute and relative criteria while keeping model states, target origins, and date weighting fixed. Explicit denominator floors are necessary for the reported zero targets, and their numerical choice changes the size of the contrast. The result is a useful benchmark lesson without requiring an unperformed algorithm comparison: the name of a model and its average error are insufficient unless the exact loss and target population accompany them.

Documentation under Gebru et al., 2021 records duplicate handling, chronological boundaries, common availability, zero-target counts, denominator conventions, units, and the daily-root calculation order. Preservation following Wilkinson et al., 2016 retains the frozen predictions and per-origin scores alongside dated contrasts. These records allow another reader to reproduce the ranking under each stated criterion. Selecting a preferred criterion or making a cross-farm operational recommendation requires additional scientific and application-specific evidence beyond this one-resource audit.

References

1. Cameron, A. Colin, and Douglas L. Miller. “A Practitioner’s Guide to Cluster-Robust Inference.” *Journal of Human Resources*, vol. 50, no. 2, 2015, pp. 317-372. https://doi.org/10.3368/jhr.50.2.317.
2. Diebold, Francis X.; Mariano, Roberto S.. (1995). Comparing Predictive Accuracy. Journal of Business & Economic Statistics. https://doi.org/10.1080/07350015.1995.10524599.
3. Perr-Sauer, Jordan; Optis, Mike; Fields, Jason; Bodini, Nicola; Lee, Joseph; Todd, Austin; Simley, Eric; Hammond, Robert; Phillips, Caleb; Lunacek, Monte; Kemper, Travis; Williams, Lindy; Craig, Anna; Agarwal, Nathan; Sheng, Shawn; Meissner, John. (2021). OpenOA: An Open-Source Codebase For Operational Analysis of Wind Farms. Journal of Open Source Software. https://doi.org/10.21105/joss.02171.
4. Roberts, David R., et al. “Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure.” *Ecography*, vol. 40, no. 8, 2017, pp. 913-929. https://doi.org/10.1111/ecog.02881.
5. Yan, Yi; Kuruoglu, Ercan E.. (2023). Fast and Robust Wind Speed Prediction Under Impulsive Noise via Adaptive Graph-Sign Diffusion. 2023 IEEE Conference on Artificial Intelligence (CAI), 302-305. https://doi.org/10.1109/CAI54212.2023.00135.
6. Gneiting, Tilmann. (2011). Making and Evaluating Point Forecasts. Journal of the American Statistical Association. https://doi.org/10.1198/jasa.2011.r10138.
7. Bergmeir, Christoph; Hyndman, Rob J.; Koo, Bonsoo. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. Computational Statistics & Data Analysis. https://doi.org/10.1016/j.csda.2017.11.003.
8. Varma, Sudhir and Richard Simon. “Bias in error estimation when using cross-validation for model selection.” *BMC Bioinformatics*, vol. 7, no. 1, 2006. https://doi.org/10.1186/1471-2105-7-91.
9. Willmott, CJ; Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research. https://doi.org/10.3354/cr030079.
10. Hodson, Timothy O.. (2022). Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. Geoscientific Model Development. https://doi.org/10.5194/gmd-15-5481-2022.
11. ENGIE / NREL OpenOA contributors, . (2026). La Haute Borne 2014-2015 SCADA and variable dictionary: captured OpenOA example archive. Local snapshot captured 7 November 2026; observed data years 2014-2015. https://github.com/NREL/OpenOA/blob/main/examples/data/la_haute_borne.zip.
12. Gneiting, Tilmann; Raftery, Adrian E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association. https://doi.org/10.1198/016214506000001437.
13. Foley, Aoife M.; Leahy, Paul G.; Marvuglia, Antonino; McKeogh, Eamon J.. (2012). Current methods and advances in forecasting of wind power generation. Renewable Energy. https://doi.org/10.1016/j.renene.2011.05.033.

14. Newey, Whitney K., and Kenneth D. West. “A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix.” *Econometrica*, vol. 55, no. 3, 1987, pp. 703-708. https://doi.org/10.2307/1913610.

15. Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; Axton, Myles; Baak, Arie; Blomberg, Niklas; Boiten, Jan-Willem; da Silva Santos, Luiz Bonino; Bourne, Philip E.; Bouwman, Jildau; Brookes, Anthony J.; Clark, Tim; Crosas, Mercè; Dillo, Ingrid; Dumon, Olivier; Edmunds, Scott; Evelo, Chris T.; Finkers, Richard; Gonzalez-Beltran, Alejandra; Gray, Alasdair J.G.; Groth, Paul; Goble, Carole; Grethe, Jeffrey S.; Heringa, Jaap; ’t Hoen, Peter A.C; Hooft, Rob; Kuhn, Tobias; Kok, Ruben; Kok, Joost; Lusher, Scott J.; Martone, Maryann E.; Mons, Albert; Packer, Abel L.; Persson, Bengt; Rocca-Serra, Philippe; Roos, Marco; van Schaik, Rene; Sansone, Susanna-Assunta; Schultes, Erik; Sengstag, Thierry; Slater, Ted; Strawn, George; Swertz, Morris A.; Thompson, Mark; van der Lei, Johan; van Mulligen, Erik; Velterop, Jan; Waagmeester, Andra; Wittenburg, Peter; Wolstencroft, Katherine; Zhao, Jun; Mons, Barend. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. https://doi.org/10.1038/sdata.2016.18.

16. Tashman, Leonard J.. (2000). Out-of-sample tests of forecasting accuracy: an analysis and review. International Journal of Forecasting. https://doi.org/10.1016/s0169-2070(00)00065-0.

17. Hyndman, Rob J. and Anne B. Koehler. “Another look at measures of forecast accuracy.” *International Journal of Forecasting*, vol. 22, no. 4, 2006, pp. 679-688. https://doi.org/10.1016/j.ijforecast.2006.03.001.

18. Makridakis, Spyros; Hibon, Michèle. (2000). The M3-Competition: results, conclusions and implications. International Journal of Forecasting. https://doi.org/10.1016/s0169-2070(00)00057-1.

19. Chai, T.; Draxler, R. R.. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? - Arguments against avoiding RMSE in the literature. Geoscientific Model Development. https://doi.org/10.5194/gmd-7-1247-2014.

20. Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. “On Calibration of Modern Neural Networks.” *Proceedings of the 34th International Conference on Machine Learning*, PMLR, vol. 70, 2017, pp. 1321-1330. https://proceedings.mlr.press/v70/guo17a.html.

21. Goodwin, Paul; Lawton, Richard. (1999). On the asymmetry of the symmetric MAPE. International Journal of Forecasting. https://doi.org/10.1016/s0169-2070(99)00007-2.

22. Benjamini, Yoav and Yosef Hochberg. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 57, no. 1, 1995, pp. 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

23. Pinson, Pierre. (2013). Wind Energy: Forecasting Challenges for Its Operational Management. Statistical Science. https://doi.org/10.1214/13-sts445.

24. Haslett, John; Raftery, Adrian E.. (1989). Space-Time Modelling with Long-Memory Dependence: Assessing Ireland's Wind Power Resource. Applied Statistics. https://doi.org/10.2307/2347679.

25. White, Halbert. “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.” *Econometrica*, vol. 48, no. 4, 1980, pp. 817-838. https://doi.org/10.2307/1912934.

26. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.

27. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.

28. West, Kenneth D.. (1996). Asymptotic Inference about Predictive Ability. Econometrica. https://doi.org/10.2307/2171956.

29. Wasserstein, Ronald L.; Lazar, Nicole A.. (2016). The ASA Statement on p -Values: Context, Process, and Purpose. The American Statistician. https://doi.org/10.1080/00031305.2016.1154108.

30. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.