Operational Evidence and Outcomes
Volume 1 · Number 1 · 2026

Robust Fitting Is Not Input Protection: A Held-Out Wind Forecasting Stress Test

Oliver Webb[1] · Joseph Bennett[2] · Evan Anderson[3]

[1] Australia · [2] Germany · [3] United Kingdom

Original publication: November 8, 2026 | Article 605

Abstract

Does fitting with a robust residual loss protect a forecast from a corrupted input? Using Irish daily wind records, we trained ordinary and Huber regressions before 1976 and evaluated 1,096 subsequent days. Huber fitting reduced clean-input mean absolute error from 3.615 to 3.591 knots. When isolated eight-standard-deviation spikes affected one lagged station value on approximately ten percent of test days, both errors rose by about 0.248 knots. The change in Huber's relative advantage was -0.00020 knots, with a 95 percent interval of [-0.01071, 0.01030]. This controlled example separates a modest forecasting advantage from evidence of additional resilience to inference-time noise.

Research draft combining historical public observations with explicitly simulated input errors. The adaptive graph-sign diffusion method is not executed here.

Robust Fitting Is Not Input Protection: A Held-Out Wind Forecasting Stress Test

Research draft combining historical public observations with explicitly simulated input errors. The adaptive graph-sign diffusion method is not executed here.

Abstract

Does fitting with a robust residual loss protect a forecast from a corrupted input? Using Irish daily wind records, we trained ordinary and Huber regressions before 1976 and evaluated 1,096 subsequent days. Huber fitting reduced clean-input mean absolute error from 3.615 to 3.591 knots. When isolated eight-standard-deviation spikes affected one lagged station value on approximately ten percent of test days, both errors rose by about 0.248 knots. The change in Huber's relative advantage was -0.00020 knots, with a 95 percent interval of [-0.01071, 0.01030]. This controlled example separates a modest forecasting advantage from evidence of additional resilience to inference-time noise.

1. Which part of a forecasting system is robust?

The word robust can describe several different protections. An estimator may reduce the influence of large training residuals, a sensor filter may suppress isolated measurement spikes, or a forecasting system may maintain accuracy when its incoming covariates change. These are related ambitions, but they operate at different points in a pipeline. A loss function used while fitting historical outcomes does not automatically constrain how strongly a fitted predictor responds to an extreme value supplied during deployment. Evaluating only average clean-data accuracy leaves that distinction unresolved.

Wind prediction under impulsive noise is the explicit topic of Fast and Robust Wind Speed Prediction Under Impulsive Noise via Adaptive Graph-Sign Diffusion (Yan and Kuruoglu, 2023). Our contribution is a small classical baseline experiment motivated by that application. We compare ordinary least squares with Huber residual fitting, using identical covariates and a chronological test period. We then perturb held-out inputs while preserving the observed wind outcomes. This exercise establishes neither the performance nor the implementation correctness of the cited graph-based method; it clarifies what a residual-robust baseline can demonstrate on its own.

The research question concerns incremental resilience: does Huber fitting reduce the additional absolute forecast error caused by isolated inference-time spikes? A secondary comparison concerns its clean-input accuracy. The distinction matters operationally because a system may have a slightly better baseline forecast while suffering nearly the same deterioration when a station reading is corrupted. Forecasting discussions that connect meteorological uncertainty with operational decisions provide a reason to keep both quantities visible (Pinson, 2013). We examine daily mean speed rather than electricity production, market loss, or turbine-control performance.

2. Robust statistics and network observations

Huber's classical work develops robust estimation through an alternative treatment of residuals (Huber, 1964). Influence-curve analysis gives another language for asking how observations affect an estimator, while least-median-of-squares regression represents a different robust regression tradition (Hampel, 1974; Rousseeuw, 1984). These sources do not imply that all robust procedures share the same breakdown behavior or response to covariate contamination. We use one conventional Huber specification and identify the location of the artificial noise explicitly, instead of interpreting the general label robust as a universal performance guarantee.

Influential observations can arise through the predictor space as well as through outcome residuals. Regression influence diagnostics make this distinction consequential (Cook, 1977). A linear prediction remains a weighted sum of its supplied covariates after training. Changing a covariate by many historical standard deviations can therefore change the prediction even when the fitted coefficients were obtained with residual reweighting. The present experiment studies that deployed calculation. It does not contaminate the training outcomes or claim to measure resistance to every form of leverage during estimation.

Meteorological stations also form a spatially related collection. Graph signal processing develops tools for data on irregular network domains (Shuman et al., 2012; Ortega et al., 2017). Those perspectives motivate considering multiple station histories, but a regression containing several station lags is not thereby a graph filter. We do not estimate an adjacency matrix, select graph edges, or apply spectral operators. All twelve available previous-day station speeds enter both baseline forecasts, allowing the comparison to isolate the fitting criterion within a clearly specified conventional design.

Several alternatives remain outside that design. Quantile regression can target conditional quantiles rather than the squared-error conditional mean (Koenker and Bassett, 1978), and probabilistic forecasts require evaluation appropriate to their reported distributions (Gneiting and Raftery, 2007). Our outputs are single-valued predictions with absolute and squared-error summaries. They do not provide calibrated prediction intervals or an optimized asymmetric operating rule. This restriction keeps the experiment interpretable while recognizing that resilience for a decision maker may require a different forecasting target.

3. Historical data and the forward test boundary

The public Irish wind dataset records daily mean speeds at twelve synoptic stations from 1961 through 1978. Its associated statistical study analyzes spatial and temporal dependence in this resource (Haslett and Raftery, 1989). The downloaded data contain 6,574 dated rows and fifteen numerical columns: three date components followed by twelve station measurements. We verified date uniqueness, parsed all records, and retained the original file hash. The accompanying StatLib description specifies knots, an essential detail because interpreting the same numbers as metres per second would change every reported magnitude.

The forecast outcome is the RPT station's daily mean speed. Predictors are the previous day's speeds at RPT, VAL, ROS, KIL, SHA, BIR, DUB, CLA, MUL, CLO, BEL, and MAL, plus sine and cosine annual terms and an intercept. Creating lags removes the first day. Training then covers 5,477 days from 2 January 1961 to 31 December 1975; testing covers 1,096 days from 1 January 1976 to 31 December 1978. The outcome station was specified as the first documented station, rather than selected for a favorable comparison.

Each lagged speed is centered and scaled using its training mean and standard deviation. The annual terms use the calendar day within a 365.25-day cycle. Neither scaling nor fitting uses test outcomes. The models remain fixed throughout the held-out period. For forecasts after the first test day, actual previous-day observations are available as lag inputs; these are sequential one-day forecasts rather than a recursive three-year forecast. Dataset documentation principles support preserving these availability and transformation assumptions with the analysis (Gebru et al., 2021).

Chronology is substantive here. Randomly mixing days can expose an evaluator to a different information pattern than prospective forecasting. Structured validation addresses dependence among observations, and work on autoregressive cross-validation identifies assumptions under which particular schemes may be

justified (Roberts et al., 2017; Bergmeir et al., 2018). We use one forward split to make the deployment boundary visible. The split is not presented as the uniquely optimal assessment design, and the historical test period cannot establish transport to contemporary instruments, station locations, or weather regimes.

4. Controlled spikes and two propositions

Both forecasts use exactly the same training matrix. Ordinary least squares minimizes squared training residuals. The second fit uses a Huber loss with tuning constant 1.345 and a maximum of 200 fitting iterations. We do not search this constant on test performance. P1 concerns the mean test absolute-error difference, defined as Huber error minus ordinary-regression error under clean inputs. Negative values favor Huber fitting. This is an accuracy comparison between the two particular fitted baselines, not a general ranking of robust and non-robust forecasting methods.

The stress test modifies only the held-out input vectors. A fixed random seed, 20261106605, generates an ordering of test days, a selected lagged-station feature for each day, and an equally signed positive or negative shift. Nominal affected-day rates are zero, one, five, and ten percent. Because days are discrete, the actual nonzero counts are 11, 55, and 110. One selected standardized lag value on each affected day is shifted by eight training standard deviations. The noise subsets are nested, permitting paired comparison as the affected set grows.

P2 asks whether Huber's relative advantage changes from clean inputs to the ten-percent scenario. We calculate a daily difference of loss differences: the Huber-minus-ordinary absolute-error difference with spikes, minus the corresponding difference without spikes. A negative mean indicates an additional relative advantage under this perturbation. Crucially, a negative P1 does not imply a negative P2. A model can forecast slightly better in both environments while losing essentially the same amount of accuracy when the input is corrupted.

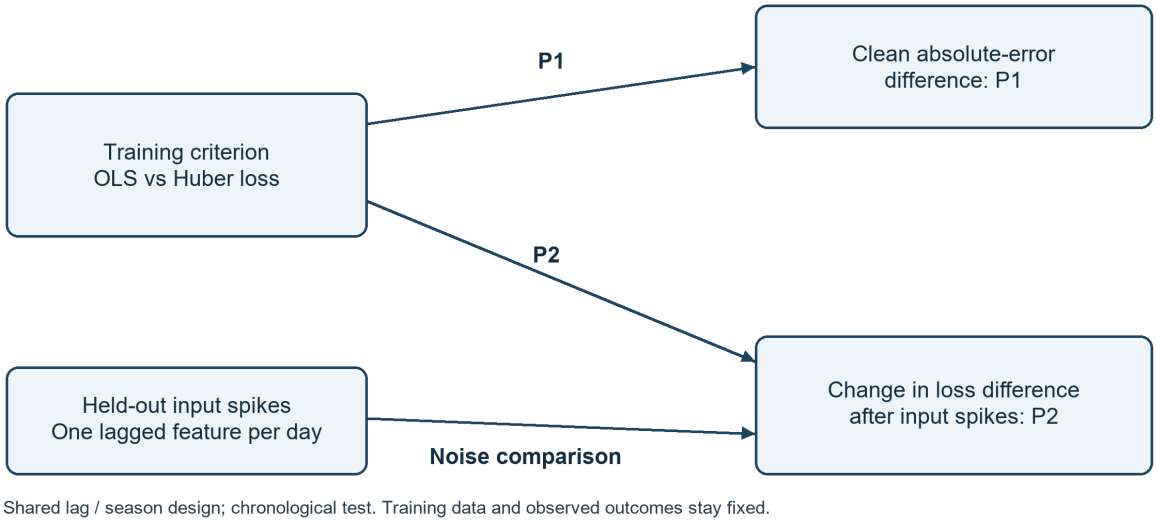


Figure 1. Research model for a paired forecasting experiment. The fitting criterion is compared under a shared lag-and-season design. P1 concerns the clean-input absolute-error difference; P2 concerns the change in that difference after inference-time spikes. The observed outcome series and fitted model coefficients remain fixed across perturbation scenarios.

The spikes are simulated errors, not historical claims about station failures. Their construction introduces an explicit form of covariate measurement error, whose consequences depend on both the error

mechanism and the model (Carroll et al., 2006). We do not truncate shifted inputs to a plausible physical range, since detecting the behavior of an unprotected numerical input is the purpose of this stress test. No evidence is supplied about the frequency or magnitude of such errors in real Irish observations, or about whether an operational quality-control system would already reject them.

5. Paired loss inference

For each scenario, we calculate mean absolute error and root mean squared error in knots. Error measures encode different penalties, so their interpretation should remain attached to the forecasting question (Hyndman and Koehler, 2006). The main inferential outcome is the daily absolute-error difference between the two forecasts. We estimate its mean through an intercept-only ordinary-least-squares equation. This coefficient is directly interpretable as an average paired loss difference; its R-squared is zero by construction, and it is not a measure of either forecasting model's explanatory fit.

Forecast-accuracy comparison requires attention to the temporal behavior of loss differences (Diebold and Mariano, 1995; West, 1996). We use a heteroskedasticity-and-autocorrelation-consistent covariance estimate with seven daily lags, finite-sample correction, and a t-based interval (Newey and West, 1987). An independence-oriented heteroskedasticity correction would address a different covariance assumption (White, 1980). The chosen weekly lag window is transparent but not guaranteed to capture every form of longer dependence in this historical series. The reported intervals are conditional on that uncertainty specification.

Time-series resampling would provide another possible sensitivity analysis, but its blocks or model assumptions require explicit choices (Bühlmann, 2002). We do not report a bootstrap that was never executed. Nor do the many test days create independent replications of the artificial noise process: each scenario uses one fixed realization. The intervals describe temporal uncertainty in the paired loss series under the fitted covariance model, while variation across alternative noise draws remains unmeasured.

Table 1. Held-out error and paired mean-loss estimates. Each scenario contains the same 1,096 test days. Parentheses contain HAC standard errors with seven daily lags; brackets contain 95 percent intervals. All error values are in knots. The last row estimates P2, comparing relative losses at ten-percent versus zero contamination. Loss-mean regressions are intercept-only with R-squared zero; forecasting covariates and training dates are shared across the two fitted systems.

Input scenario	OLS MAE	Huber MAE	Difference (HAC SE)	95% CI
0% noisy input days	3.6153	3.5913	-0.02404 (0.00717)	[-0.03811, -0.00998]
1% noisy input days	3.6433	3.6195	-0.02380 (0.00729)	[-0.03810, -0.00950]
5% noisy input days	3.7247	3.7021	-0.02263 (0.00780)	[-0.03794, -0.00732]
10% noisy input days	3.8637	3.8394	-0.02425 (0.00851)	[-0.04094, -0.00755]
P2: ten-percent minus clean	-	-	-0.00020 (0.00536)	[-0.01071, 0.01030]

6. A small baseline advantage, little additional resilience

With unmodified inputs, ordinary regression has mean absolute error 3.6153 knots and Huber regression 3.5913. Their paired difference is -0.02404 knots, with interval [-0.03811, -0.00998]. The result supports a modest average advantage for this Huber fit during the held-out period. It does not establish that the difference is large enough to change an operational decision. No cost function for reserve scheduling, sensor maintenance, or turbine protection was supplied, so we retain the physical error scale instead of attaching an invented monetary value.

When 110 test days receive a spike, the corresponding mean absolute errors become 3.8637 and 3.8394 knots. Both increase by approximately 0.248 knots from their clean values. Huber's relative difference remains -0.02425, with interval [-0.04094, -0.00755]. Thus a table showing only its better error in the noisy scenario could suggest resilience without establishing it. The appropriate incremental comparison asks whether that advantage exceeds the one already present before corruption.

P2's estimated difference of differences is -0.00020 knots, with standard error 0.00536 and interval [-0.01071, 0.01030]. The result does not identify a clear extra protection from residual-robust fitting under this input mechanism. Intermediate one- and five-percent scenarios show the same broad pattern: both systems deteriorate while their relative absolute-error difference stays small. We preserve all rates rather than selecting the scenario with the largest favorable gap. Root mean squared errors also increase, but they are secondary descriptive summaries rather than a separately optimized conclusion.

7. Implications and limitations

The operational lesson concerns pipeline design. Training loss and incoming-data protection should be specified separately. A linear predictor remains exposed to numerical spikes unless its coefficients or input processing provide relevant protection. This experiment leaves room for clipping, robust covariate transformations, graph-based processing, and model-specific defenses, but evaluates none of them. A future comparison should preserve the same forward test boundary and describe precisely where each defense acts, rather than comparing systems whose fitting and input-cleaning choices change together.

Pooling all twelve station outcomes would require additional design decisions. Repeated daily losses within stations and common weather across stations create dependencies; treating every station-day as independent would be difficult to defend. Cluster-based and multilevel approaches offer relevant frameworks when their assumptions and target population are specified (Cameron and Miller, 2015; Laird and Ware, 1982; Bates et al., 2015). Our single-outcome analysis avoids claiming multi-station replication from a design in which stations are only predictors. Its conclusion remains specific to RPT and the historical split used here.

Several further boundaries protect interpretation. No missing-value imputation was needed in this downloaded matrix, so the experiment provides no evidence about missing-sensor recovery (Little and Rubin, 2002). Predictor quality-control effects in real deployments would require a separate intervention and an appropriate causal design; observational adjustment or matching alone would not make the current simulated spikes representative of real failures (Greenland et al., 1999; Stuart, 2010). The deterministic comparison identifies what the two fitted systems do to these altered vectors, rather than an externally generalizable effect of a field maintenance policy.

An expanded study could vary stations, thresholds, noise amplitudes, training windows, and loss functions. Model selection on the evaluation period would risk optimistic reporting (Varma and Simon, 2006), while inferential screening over many variants raises multiplicity concerns (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001). We report the declared small scenario set, including the null incremental-resilience estimate. Scientific confidence would be strengthened by independent periods, alternative noise realizations, and a documented operational error distribution; those are proposed extensions rather than completed validations.

8. Conclusion

Huber residual fitting delivered slightly lower held-out wind-speed error than ordinary regression in this historical example, but its relative advantage did not clearly expand when lagged inputs received isolated artificial spikes. The relevant resilience contrast was close to zero and imprecisely signed. Keeping clean accuracy and noise-induced deterioration separate prevents a favorable baseline comparison from becoming an unsupported robustness claim. Input protection requires its own mechanism and evidence, even when the estimation procedure already carries a robust-statistics label.

References

1. Diebold, Francis X.; Mariano, Roberto S.. (1995). Comparing Predictive Accuracy. Journal of Business & Economic Statistics. https://doi.org/10.1080/07350015.1995.10524599.
2. Ortega, Antonio; Frossard, Pascal; Kovačević, Jelena; Moura, José M. F.; Vandergheynst, Pierre. (2017). Graph Signal Processing: Overview, Challenges and Applications. arXiv:1712.00468. https://arxiv.org/abs/1712.00468.
3. Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. “Fitting Linear Mixed-Effects Models Using lme4.” *Journal of Statistical Software*, vol. 67, no. 1, 2015. https://doi.org/10.18637/jss.v067.i01.
4. 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.
5. Huber, Peter J.. (1964). Robust Estimation of a Location Parameter. The Annals of Mathematical Statistics. https://doi.org/10.1214/aoms/1177703732.
6. Little, Roderick J. A., and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2002. https://doi.org/10.1002/9781119013563.
7. 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.
8. Koenker, Roger; Bassett, Gilbert. (1978). Regression Quantiles. Econometrica. https://doi.org/10.2307/1913643.
9. 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.
10. Stuart, Elizabeth A. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science*, vol. 25, no. 1, 2010, pp. 1-21. https://doi.org/10.1214/09-sts313.
11. Greenland, Sander, Judea Pearl, and James M. Robins. “Causal Diagrams for Epidemiologic Research.” *Epidemiology*, vol. 10, no. 1, 1999, pp. 37-48. https://doi.org/10.1097/00001648-199901000-00008.
12. 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.
13. Pinson, Pierre. (2013). Wind Energy: Forecasting Challenges for Its Operational Management. Statistical Science. https://doi.org/10.1214/13-sts445.
14. Benjamini, Yoav and Daniel Yekutieli. “The control of the false discovery rate in multiple testing under dependency.” *The Annals of Statistics*, vol. 29, no. 4, 2001, pp. 1165-1188. https://doi.org/10.1214/aos/1013699998.
15. Rousseeuw, Peter J.. (1984). Least Median of Squares Regression. Journal of the American Statistical Association. https://doi.org/10.1080/01621459.1984.10477105.
16. 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.
17. 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.
18. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.
19. West, Kenneth D.. (1996). Asymptotic Inference about Predictive Ability. Econometrica. https://doi.org/10.2307/2171956.
20. Shuman, David I; Narang, Sunil K.; Frossard, Pascal; Ortega, Antonio; Vandergheynst, Pierre. (2012). The Emerging Field of Signal Processing on Graphs: Extending High-Dimensional Data Analysis to Networks and Other Irregular Domains. arXiv:1211.0053. https://arxiv.org/abs/1211.0053.

21. Hampel, Frank R.. (1974). The Influence Curve and its Role in Robust Estimation. Journal of the American Statistical Association. https://doi.org/10.1080/01621459.1974.10482962.

22. Bühlmann, Peter. (2002). Bootstraps for Time Series. Statistical Science. https://doi.org/10.1214/ss/1023798998.

23. 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.

24. 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.

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. 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.

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

29. 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.

30. Laird, Nan M. and James H. Ware. “Random-Effects Models for Longitudinal Data.” *Biometrics*, vol. 38, no. 4, 1982, pp. 963. https://doi.org/10.2307/2529876.