Annals of Science and Inquiry
Vol. 2 No. 6 (2026)

Sharing Short-History Coefficients in Six-Hour Forecasts of Restored Apollo Temperatures

Patrick Warren[1] · Spencer Allen[2] · Gavin Bell[3]

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

Published: June 17, 2026 | Article 852

Abstract

Ten Apollo thermometer channels share instrumental ancestry, but their short-history forecast coefficients need not be interchangeable. Using earlier April–May records for fitting and June 1975 for evaluation, we compared persistence with channel-specific and shared-slope six-hour increment forecasts. Mean absolute errors were 33.522, 28.186 and 32.820 mK, respectively. Sharing increased error over separate fits by 4.634 mK, with a conditional calendar-HAC working interval of 1.679–7.589 mK. The shared model also worsened two channels relative to persistence. These comparisons concern restored recorded temperatures under fixed observation-support rules; they establish neither independent sensor accuracy nor an operational lunar forecasting system.

Sharing Short-History Coefficients in Six-Hour Forecasts of Restored Apollo Temperatures

Abstract

Ten Apollo thermometer channels share instrumental ancestry, but their short-history forecast coefficients need not be interchangeable. Using earlier April-May records for fitting and June 1975 for evaluation, we compared persistence with channel-specific and shared-slope six-hour increment forecasts. Mean absolute errors were 33.522, 28.186 and 32.820 mK, respectively. Sharing increased error over separate fits by 4.634 mK, with a conditional calendar-HAC working interval of 1.679-7.589 mK. The shared model also worsened two channels relative to persistence. These comparisons concern restored recorded temperatures under fixed observation-support rules; they establish neither independent sensor accuracy nor an operational lunar forecasting system.

A common instrument family does not determine a common forecast

Forecasting several temperature channels presents a choice before any elaborate model is considered. One can estimate a response to recent movement separately for each channel, or require that response to be shared. Sharing reduces the number of fitted coefficients and allows information from more records to influence each forecast. It also imposes a restriction: the same recent increment must imply the same predicted subsequent increment after channel intercepts are allowed. Whether that restriction helps is a question about later predictive error, rather than a consequence of common units or instrument names.

The distinction between local and global forecasting has a formal literature. Montero-Manso and Hyndman describe learning separately from individual series versus learning across a group, including global linear methods (Montero-Manso and Hyndman, 2021). Their analysis does not require every particular shared linear specification to dominate separate fits. Characteristics such as parameter sharing and information access are more informative here than classifying the forecasts as statistical or machine learning (Januschowski et al., 2020). DeepAR illustrates a substantially richer approach that trains an autoregressive recurrent model across related series (Salinas et al., 2020). We use it only to locate the broader idea of learning across series; no neural architecture or probabilistic DeepAR forecast is implemented.

Our research question is precise: for a fixed six-hour recorded-temperature target, how does sharing two history-response coefficients across ten restored Apollo channels change later error relative to channel-specific coefficients and persistence? Proposition P1 is that fitting a temporal response need not improve on carrying forward the latest value. Proposition P2 is that the consequences of sharing can differ across channels even when they have the same units and instrument family. Neither proposition assigns a causal role to a coefficient or predicts that every channel will follow the aggregate ranking.

The empirical contribution is a controlled comparison of parameter constraints using the same histories, targets and scored observations. It does not add a second history feature to an otherwise different score, infer a periodic component, or assess a physical heat-transfer law. A later chronological interval supplies the evidence needed to distinguish fit from forecast. All numerical comparisons below follow a protocol saved after source inspection and before model fitting, with no selection based on June errors.

What the restored temperatures represent

The source consists of the six original calibrated gradient-bridge products for day-of-year 092-181 in 1975 from the Apollo 15 and Apollo 17 ARCSAV collections (Nagihara and Nakamura, 2018a) (Nagihara and Nakamura, 2018b). The acquired ASCII products contain 99,414 rows. A row records an Earth-received UTC timestamp, upper and lower resistance-temperature-detector values, their bridge average, and high- and low-gain temperature differences. All temperature columns use kelvins. We retain the provider's explicit upper and lower temperatures, rather than construct a new reduction from rounded difference fields.

The two detector values within a bridge derive from a shared average and a selected differential measurement. Their common inputs matter: ten channels are not ten independent instrument experiments. The later concatenated archive documents the thermometer nomenclature, quality annotations, historical depth choices and saturated bridges (St. Clair et al., 2019). Its associated publication discusses improvements to the Apollo heat-flow archive (St. Clair et al., 2020). We use that documentation to interpret the original calibrated products, while preserving their own timestamp field as the primary clock.

One bridge cannot support the intended subsurface comparison. Apollo 15 probe 2 TG21 includes an upper detector above the surface and is documented as frequently saturated. Its 17,411 rows remain in the source inventory but neither channel is fitted or scored. The other five bridges provide ten channels in four probes and two mission sites. Source-marked temperature or differential outliers exclude another 77 paired bridge records. No residual-based filter or new temperature threshold is applied. Finite negative differences are retained, and an absent observation never becomes a zero temperature or zero error.

The original UTC and later clean-format clock representations differ by constant offsets of approximately 23 seconds. We do not interpret that discrepancy as an established correction. Quality flags are linked after separately ordering each representation and verifying a complete correspondence using bridge averages and signed original difference fields. Original record identifiers remain attached to every selected observation. The source has three row-order reversals in one Apollo 17 product, but no repeated timestamps; chronological processing therefore sorts actual times without changing them.

The recorded quantity must stay distinct from a stronger physical interpretation. Metrological vocabulary distinguishes a specified quantity from the wider object one might wish to characterize (JCGM, 2012). Here the target is a later provider-calibrated recorded temperature. A predictive-loss interval is also different from a complete measurement-uncertainty budget (JCGM, 2008). Printed 0.001 K resolution and documented differential measurement precision do not provide that complete budget. Heater states, operating disturbances and independent ground truth are not established by these six products.

These qualifications affect fitness for forecasting reuse, rather than merely the cleanliness of a spreadsheet. Data quality includes the consumer's intended use beyond numerical accuracy alone (Wang and Strong, 1996). We therefore retain a source inventory with product versions, byte hashes, field roles, quality exclusions and acquisition information. Such documentation serves the provenance and intended-use questions emphasized by dataset documentation work (Gebru et al., 2021). The PDS FAQ documents free access, TSPA classification and export as public open literature with appropriate citation; that statement is not a blanket copyright license for archived publications. Nor does source documentation certify that retrospective archive quality flags would have been available online at forecast time.

Defining a future endpoint without filling the gaps

The analysis window starts at 00:00 UTC on 2 April and ends immediately before 1 July 1975. Typical consecutive observations are about 434.71 seconds apart, but the largest source gap is approximately five days. Treating the next stored row as a fixed forecast horizon would therefore mix very different elapsed times. We instead place candidate origins at exact UTC hours, irrespective of how many source rows happen to occur nearby.

For an origin t, four separate queries retrieve the latest unflagged observation at or before t-2 hours, t-1 hour, t and t+6 hours. Every selected value must be no more than fifteen minutes old relative to its own query. Consecutive retained-source observations across the selected two-hour history must also be at most fifteen minutes apart. The latter rule prevents endpoint values from concealing a long interruption within the short history. No interpolation, nearest future match or filled hourly series is created.

The future endpoint supplies a recorded label for fitting earlier examples or scoring later forecasts; it never becomes a predictor. Complete observation of the intervening six-hour path is unnecessary for this particular endpoint target, and we do not claim to have observed such a continuous path. If its endpoint is stale or missing, a candidate remains unscored. The resulting observed current-to-target separations in the principal test range from 20,866.044 to 22,170.532 seconds. Thus six hours denotes the nominal UTC query horizon; it does not falsely assert that every pair of selected source observations is exactly six hours apart.

Features are the two consecutive hourly-coordinate increments, x1 = T(t)-T(t-1 hour) and x2 = T(t-1 hour)-T(t-2 hours), using those as-of values. They are temperature differences in kelvins, not derivatives calculated with an invented exact sampling interval. Only this same channel's past readings enter its forecast. The shared model transfers fitted coefficients; it does not consume another channel's simultaneous value or any future reading as a predictor.

Training labels must have their nominal six-hour endpoint strictly before 1 June. Origins in the final six hours of May are consequently withheld from fitting even when their past observations are supported. June origins form the test interval, provided their nominal endpoint precedes 1 July. Past observations from May can legitimately support an early June origin. Model coefficients remain fixed throughout evaluation. This distinction between fit/test division and coefficient updating is central to out-of-sample forecast assessment (Tashman, 2000).

Chronological evaluation answers a particular extrapolation question. Validation research warns that temporal and hierarchical structure can change the meaning of predictive-error estimates (Roberts et al., 2017). Comparisons of cross-validation and last-block procedures also depend on the evolving properties of the series (Bergmeir and Benítez, 2012). Empirical evaluations of time-series performance-estimation methods further motivate making the evaluation design explicit (Cerqueira et al., 2020). We use one fixed later month; it supplies no evidence about repeating this result across years, mission sites or newly installed instruments.

Three forecasts and an explicit sharing restriction

Persistence predicts zero six-hour change, so its predicted endpoint is the current as-of temperature. The local fitted forecast predicts the increment as alpha_c + beta1_c x1 + beta2_c x2, where c identifies a channel. The shared-response forecast uses alpha_c + beta1 x1 + beta2 x2. It retains a different intercept for every channel while restricting both slopes to be common. The local collection has thirty coefficients; the shared fit has twelve.

These are direct fixed-horizon increment regressions. They do not recursively apply a one-hour equation six times, estimate a stationary autoregressive likelihood, or identify physical dynamical parameters. Multi-step forecasting research distinguishes such strategy choices and their consequences (Ben Taieb et al., 2012). The specification here deliberately keeps the same two predictors and horizon in both fitted alternatives. Any difference between their later scores concerns the fitted sharing constraint and its resulting predictions, rather than different available input information.

Training weights first equalize dates within each channel, then equalize the ten channels. A row on a channel-date receives weight 1/(D_c n_cd), where D_c is that channel's number of training dates and n_cd its retained origins on that date. Every channel therefore contributes total weight one to the shared squared-error objective. Local regressions use the same relative date weights. Weighting is a definition of the represented comparison (Solon et al., 2015); these are neither inverse calibration-variance weights nor inclusion probabilities for a lunar population.

All ten local designs have full rank three, and the shared design has full rank twelve. Weighted design condition numbers range from 26.60 to 43.94 locally and are 11.51 for the shared model. The common slopes are -0.131841 and 0.108709. Their signs and sizes describe a fitted six-hour increment rule conditional on this training objective. They are not causal responses to changing a thermometer's recent history. No penalty, lag order, depth subset or alternative horizon is selected using test performance.

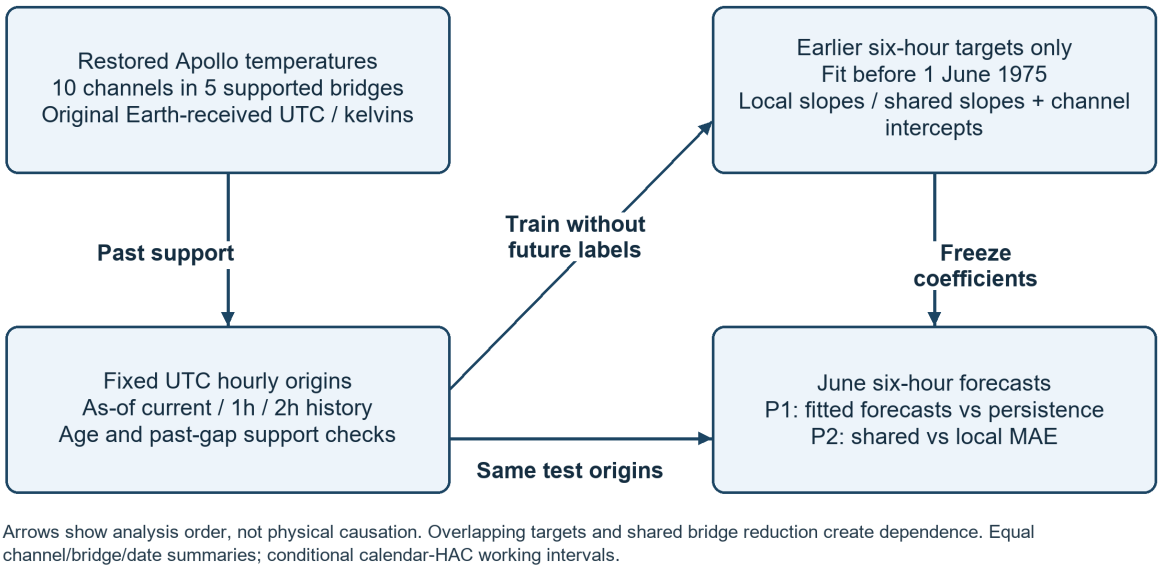

Figure 1. The same supported past observations feed the local and shared-response forecasts. Earlier labels fit the coefficients; later recorded endpoints score the frozen predictions. P1 compares fitted forecasts with persistence, and P2 compares the coefficient-sharing restriction. Arrows describe analysis order.

Loss aggregation and the limits of its working interval

Absolute error in kelvins is the primary loss. We average within channel and origin date, average the two channels within each bridge, then average the five bridges and the common dates. Root mean squared error uses the same weights before taking the square root. Dimensional accuracy measures have different sensitivity to large errors, making it useful to retain both while specifying which is primary (Hyndman and Koehler, 2006). Results are displayed in millikelvins for readability, using exactly 1,000 mK per K.

The source generates 25,920 candidate channel-origins, including the excluded bridge's audit rows. There are 11,684 supported training and 6,598 supported test channel-origins in the five valid bridges. Every

June date has at least one supported forecast in every bridge, so all thirty dates enter the principal daily comparison. No supported June forecast lies outside that common-date set. Earlier Apollo 15 training channels have sixty observed dates, while Apollo 17 channels have forty-nine. Equal-channel weighting prevents the denser training coverage from silently assigning more aggregate weight to Apollo 15.

We compare forecasts through paired daily loss differences, consistent with the loss-differential perspective of forecast evaluation (Diebold and Mariano, 1995). The primary contrast is shared minus local MAE; local minus persistence and shared minus persistence are retained secondary contrasts. Positive values mean greater error for the first forecast. These are comparisons of the executed, fixed predictions. Work on conditional predictive ability considers broader inferential questions for possibly misspecified forecasting procedures (Giacomini and White, 2006). Our working intervals do not implement that full testing framework.

For each contrast, a Bartlett HAC calculation uses seven actual calendar-day lags and a thirty-over-twenty-nine finite-sample multiplier. The covariance construction follows the heteroskedasticity-and-autocorrelation-consistent principle (Newey and West, 1987). Calendar separation determines which daily pairs contribute; missing days would not become adjacent observations or zero losses. In the realized complete June set, lags one through seven contain twenty-nine through twenty-three pairs. Bandwidth and kernel remain working choices, as the broader HAC literature makes clear (Andrews, 1991).

The displayed normal intervals assume a sufficiently representative, weakly dependent daily loss process over this short interval. That assumption is not established by the hourly record count. Six-hour targets overlap, days can remain dependent beyond seven lags, and the two channels of a bridge share reduction inputs. Training-fit uncertainty, calibration uncertainty and variation across hypothetical lunar sites are excluded. Nested-model forecast tests can require estimation-sensitive asymptotics (Clark and McCracken, 2001); we supply no nested-model p-values or universal equal-accuracy test.

The shared fit preserves little of the local improvement

Table 1 gives the aggregate scores. The local model has the smallest error under both losses: MAE is 28.186 mK and RMSE is 40.805 mK. Persistence gives 33.522 and 51.851 mK. The shared-response model gives 32.820 and 51.124 mK, remaining much closer to persistence than to the local model in this evaluation.

Table 1. Frozen six-hour forecast errors on the same 6,598 supported June channel-origins, summarized over thirty dates, ten channels and five bridges. Rows are model descriptions, not independent experimental replications.

Forecast	MAE (mK)	RMSE (mK)
persistence	33.522	51.851
local	28.186	40.805
shared	32.820	51.124

All channels appear in Table 2. The local model improves the observed MAE over persistence in each of these ten channel summaries. The shared model does not: both Apollo 17 probe 1 channels worsen, from 21.851 to 22.281 mK and from 21.927 to 22.378 mK. Its overall gain over persistence therefore cannot stand in for a benefit in every constituent series. The local/shared ranking is also uneven in magnitude. Apollo 15 TG22A differs by only about 0.081 mK, whereas the shallow TG11A channel differs by about 24.370 mK.

Table 2. Channel-specific date-balanced June MAE. A/B rows within a bridge share origin support and source reduction inputs. Fit/June counts denote overlapping forecast records; channel means have no independent-sensor uncertainty interval.

Channel	Fit rows	June rows	Persistence MAE (mK)	Local MAE (mK)	Shared MAE (mK)
A15P1TG11A	1297	658	97.961	71.998	96.368
A15P1TG11B	1297	658	25.561	21.027	23.856
A15P1TG12A	1299	656	25.856	21.080	25.012
A15P1TG12B	1299	656	25.936	21.118	25.051
A15P2TG22A	1288	648	40.089	39.547	39.628
A15P2TG22B	1288	648	24.724	20.490	23.743
A17P1TG11A	983	672	21.851	20.601	22.281
A17P1TG11B	983	672	21.927	20.619	22.378
A17P2TG21A	975	665	25.669	22.552	24.936
A17P2TG21B	975	665	25.645	22.824	24.943

The primary shared-minus-local contrast is 4.634 mK, with a working HAC standard error of 1.507 mK and interval from 1.679 to 7.589 mK. The two secondary contrasts are -5.336 mK for local minus persistence and -0.702 mK for shared minus persistence. Table 3 keeps their different magnitudes visible. Its pointwise intervals are conditional summaries, without a family-wide discovery claim. A small average numerical gain does not by itself establish an operationally useful temperature improvement.

Table 3. Paired daily MAE contrasts, thirty dates. Working intervals use fixed seven-calendar-day Bartlett HAC and exclude training/calibration uncertainty. The last column deletes one bridge at a time from the same dates and fixed predictions; its five-value range is deterministic, not a confidence interval.

MAE contrast	Effect (mK)	HAC7 SE (mK)	Working 95% lower	Working 95% upper	Finite bridge-deletion range (mK)
shared minus local	4.634	1.507	1.679	7.589	2.393 to 5.376
local minus persistence	-5.336	1.312	-7.907	-2.765	-6.351 to -2.858
shared minus persistence	-0.702	0.273	-1.237	-0.167	-0.988 to -0.466

The primary difference remains positive after deleting any one bridge, ranging from 2.393 to 5.376 mK. Removing Apollo 15 probe 1 TG11 gives the smallest value, showing that this bridge contributes strongly to the aggregate difference. Yet its deletion does not reverse that particular ordering. This exercise describes contribution concentration among the five observed bridges. It neither refits the forecasting models nor treats the five exclusions as independent experiments.

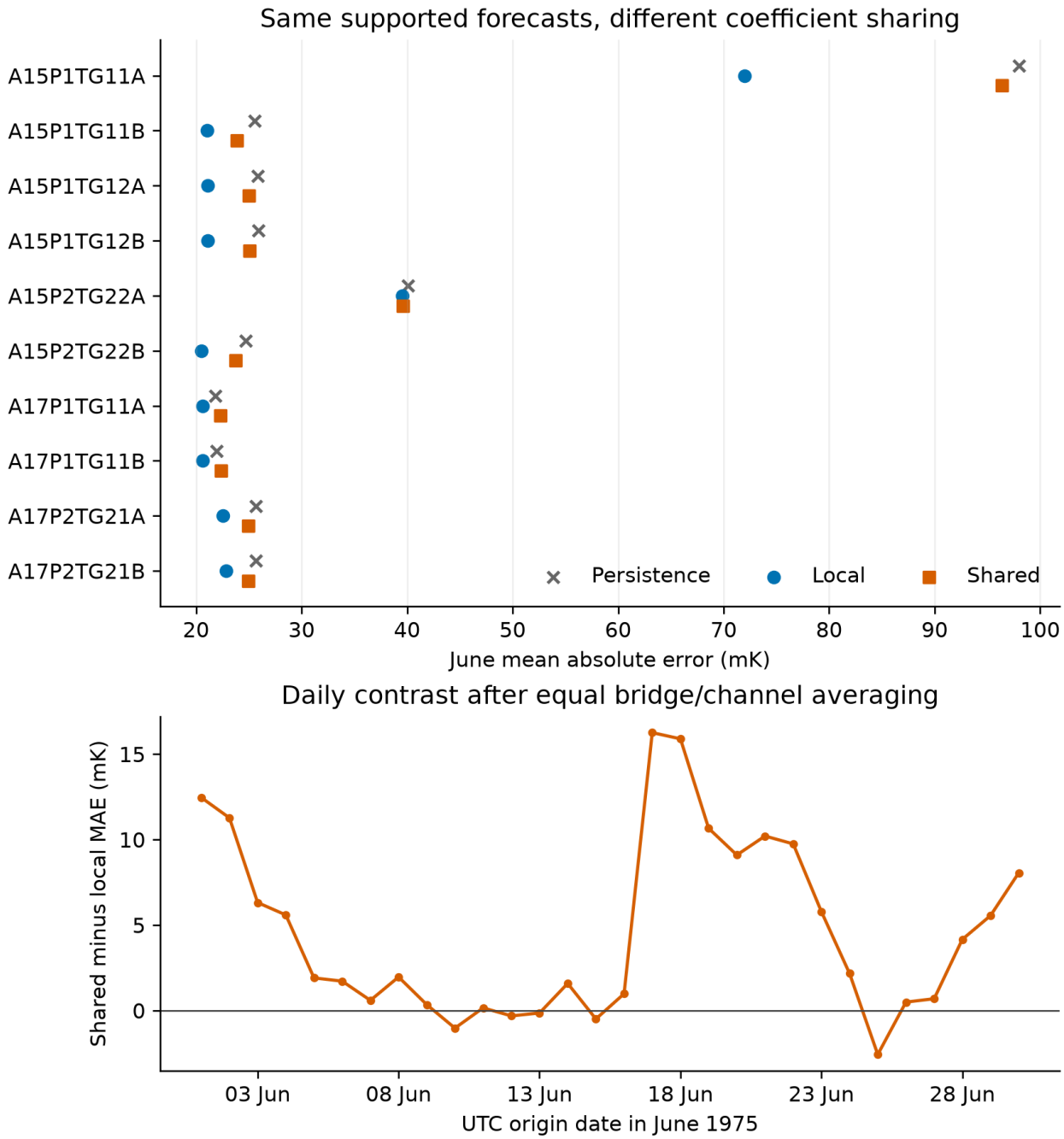

Figure 2. Upper panel: all ten channel means, with marker shapes distinguishing the three forecasts. Lower panel: every observed June daily shared-minus-local contrast after equal bridge/channel averaging. Points are computed losses; lines join consecutive observed dates without smoothing. Neither panel presents independent-channel confidence bounds.

The daily contrast is not uniformly positive. The plotted losses include days when sharing performs slightly better and stretches when its disadvantage is larger. Thus the mean ranking does not describe a constant penalty attached to each forecast origin. Equally, the fitted common slopes do not identify why the histories differ. Shared calibration inputs, quantization, physical depth differences and unrecorded operating changes are plausible contributors to recorded behavior, but this design separates none of them causally.

Implications for the use of lunar thermal information

The result supports evaluating a proposed sharing restriction rather than choosing it from instrument similarity alone. The shared fit has fewer coefficients and the same inputs, but its improvement over persistence is small in aggregate and changes sign in two channel summaries. Conversely, the local model's better observed scores do not justify unexamined complexity in general. This experiment compares only one very small direct increment specification under one split. More lags, recursive equations, regularization or external thermal inputs were not tested.

Lunar-Bench supplies a useful boundary for this evidence. Its methods evaluate task-oriented language-model reasoning, and its actual Table 10 presents a sampling scenario with specified regolith/tool conditions and a thermal scraper activated at 50°C (Xiao et al., 2026). Reasoning from supplied scenario conditions is a different task from establishing the prospective quality of a measured thermal input. The present forecast comparison concerns the latter measurement stream at Apollo sites. It neither reruns that benchmark nor validates its scraper choice, activation setpoint, model score or South Pole-Aitken environment. A forecast of a historical subsurface channel cannot be substituted for an unmeasured device temperature or an engineering safety limit.

Several further limits follow directly from the design. The training period contains unequal date coverage across missions, the test covers only June, and supported endpoints select which later values can be scored. The error estimates say nothing about performance through missing endpoints. Retrospective quality flags also make the study a forecast reconstruction under a declared archive-processing rule, rather than a demonstrated live measurement pipeline. No source-status filter can replace information about unavailable heater or operating states.

Numerical verification addresses a narrower issue and is nevertheless necessary. An independent calculation rereads all original source rows and quality correspondences, advances through time with a forward scan instead of binary lookup, and solves scalar weighted moments in fifty-digit decimal arithmetic. Channel-centering reduces its shared-coefficient check to a separate two-by-two system. All 25,920 candidate statuses and selected source identities agree. Maximum coefficient and endpoint-prediction differences are below 3.6×10^{-15} and 1.2×10^{-13} K, respectively; daily scores, covariance summaries and bridge-deletion contrasts also reproduce. This is an implementation check, not independent confirmation of sensor truth.

The actual workflow uses NumPy for array calculations and weighted least squares (Harris et al., 2020), pandas for timestamp/tabular operations (McKinney, 2010), SciPy's normal quantile for checking the working interval endpoints (Virtanen et al., 2020), and Matplotlib for the numerical figure (Hunter, 2007). Scientific Agent Skills informed the local writing, statistical reporting and figure-review workflow (Kassis et al., 2026); it supplies procedural guidance rather than evidence for the numerical result. Source hashes, the frozen protocol, candidates and independent calculations remain linked, supporting computational reproduction in the sense distinguished from scientific replication by Peng (Peng, 2011). Accountable human source verification and final approval remain pending.

Conclusion

For these supported June records, separate history-response coefficients improve the six-hour increment forecast more than the declared shared-slope restriction. Sharing raises MAE over local fitting by 4.634 mK under equal date, bridge and channel weighting, and the ordering survives each fixed-prediction bridge deletion. The shared model's much smaller average advantage over persistence does not hold in every channel. These findings answer a specific question about parameter sharing in restored recorded-temperature forecasts. They leave physical causation, complete measurement uncertainty, live deployment and wider lunar generalization unresolved.

References

1. Nagihara, Seiichi; Nakamura, Yosio. (2018b). Apollo 17 ALSEP ARCSAV Heat Flow Experiment Calibrated Gradient Bridge Temperatures Bundle,version1.0. NASA Planetary Data System. Edited by D.R.Williams and S.A.McLaughlin. https://doi.org/10.17189/1518451

2. Virtanen, Pauli; Gommers, Ralf; Oliphant, Travis E.; Haberland, Matt; Reddy, Tyler; Cournapeau, David; Burovski, Evgeni; Peterson, Pearu; Weckesser, Warren; Bright, Jonathan; van der Walt, Stéfan J.; Brett, Matthew; Wilson, Joshua; Millman, K. Jarrod; Mayorov, Nikolay; Nelson, Andrew R. J.; Jones, Eric; Kern, Robert; Larson, Eric; Carey, C J; Polat, İlhan; Feng, Yu; Moore, Eric W.; VanderPlas, Jake; Laxalde, Denis; Perktold, Josef; Cimrman, Robert; Henriksen, Ian; Quintero, E. A.; Harris, Charles R.; Archibald, Anne M.; Ribeiro, Antônio H.; Pedregosa, Fabian; van Mulbregt, Paul; SciPy 1.0 Contributors; Vijaykumar, Aditya; Bardelli, Alessandro Pietro; Rothberg, Alex; Hilboll, Andreas; Kloeckner, Andreas; Scopatz, Anthony; Lee, Antony; Rokem, Ariel; Woods, C. Nathan; Fulton, Chad; Masson, Charles; Häggström, Christian; Fitzgerald, Clark; Nicholson, David A.; Hagen, David R.; Pasechnik, Dmitrii V.; Olivetti, Emanuele; Martin, Eric; Wieser, Eric; Silva, Fabrice; Lenders, Felix; Wilhelm, Florian; Young, G.; Price, Gavin A.; Ingold, Gert-Ludwig; Allen, Gregory E.; Lee, Gregory R.; Audren, Hervé; Probst, Irvin; Dietrich, Jörg P.; Silterra, Jacob; Webber, James T; Slavič, Janko; Nothman, Joel; Buchner, Johannes; Kulick, Johannes; Schönberger, Johannes L.; de Miranda Cardoso, José Vinícius; Reimer, Joscha; Harrington, Joseph; Rodríguez, Juan Luis Cano; Nunez-Iglesias, Juan; Kuczynski, Justin; Tritz, Kevin; Thoma, Martin; Newville, Matthew; Kümmerer, Matthias; Bolingbroke, Maximilian; Tartre, Michael; Pak, Mikhail; Smith, Nathaniel J.; Nowaczyk, Nikolai; Shebanov, Nikolay; Pavlyk, Oleksandr; Brodtkorb, Per A.; Lee, Perry; McGibbon, Robert T.; Feldbauer, Roman; Lewis, Sam; Tygier, Sam; Sievert, Scott; Vigna, Sebastiano; Peterson, Stefan; More, Surhud; Pudlik, Tadeusz; Oshima, Takuya; Pingel, Thomas J.; Robitaille, Thomas P.; Spura, Thomas; Jones, Thouis R.; Cera, Tim; Leslie, Tim; Zito, Tiziano; Krauss, Tom; Upadhyay, Utkarsh; Halchenko, Yaroslav O.; Vázquez-Baeza, Yoshiki. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods. https://doi.org/10.1038/s41592-019-0686-2.

3. Harris, Charles R.; Millman, K. Jarrod; van der Walt, Stéfan J.; Gommers, Ralf; Virtanen, Pauli; Cournapeau, David; Wieser, Eric; Taylor, Julian; Berg, Sebastian; Smith, Nathaniel J.; Kern, Robert; Picus, Matti; Hoyer, Stephan; van Kerkwijk, Marten H.; Brett, Matthew; Haldane, Allan; del Río, Jaime Fernández; Wiebe, Mark; Peterson, Pearu; Gérard-Marchant, Pierre; Sheppard, Kevin; Reddy, Tyler; Weckesser, Warren; Abbasi, Hameer; Gohlke, Christoph; Oliphant, Travis E.. (2020). Array programming with NumPy. Nature. https://doi.org/10.1038/s41586-020-2649-2.

4. Giacomini, Raffaella; White, Halbert. (2006). Tests of Conditional Predictive Ability. Econometrica. https://doi.org/10.1111/j.1468-0262.2006.00718.x

5. Januschowski, Tim; Gasthaus, Jan; Wang, Yuyang; Salinas, David; Flunkert, Valentin; Bohlke-Schneider, Michael; Callot, Laurent. (2020). Criteria for classifying forecasting methods. International Journal of Forecasting. https://doi.org/10.1016/j.ijforecast.2019.05.008

6. McKinney, Wes. (2010). Data Structures for Statistical Computing in Python. Proceedings of the Python in Science Conference. https://doi.org/10.25080/majora-92bf1922-00a

7. Peng, Roger D.. (2011). Reproducible Research in Computational Science. Science. https://doi.org/10.1126/science.1213847

8. Joint Committee for Guides in Metrology. (2008). Evaluation of measurement data - Guide to the expression of uncertainty in measurement (JCGM100:2008). Bureau International des Poids et Mesures. https://doi.org/10.59161/JCGM100-2008E

9. Joint Committee for Guides in Metrology. (2012). International vocabulary of metrology - Basic and general concepts and associated terms (VIM), third edition (JCGM200:2012). Bureau International des Poids et Mesures. https://doi.org/10.59161/JCGM200-2012

10. Wang, Richard Y.; Strong, Diane M.. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems. https://doi.org/10.1080/07421222.1996.11518099

11. Cerqueira, Vitor; Torgo, Luis; Mozetič, Igor. (2020). Evaluating time series forecasting models: an empirical study on performance estimation methods. Machine Learning. https://doi.org/10.1007/s10994-020-05910-7

12. Roberts, David R.; Bahn, Volker; Ciuti, Simone; Boyce, Mark S.; Elith, Jane; Guillera-Arroita, Gurutzeta; Hauenstein, Severin; Lahoz-Monfort, José J.; Schröder, Boris; Thuiller, Wilfried; Warton, David I.; Wintle, Brendan A.; Hartig, Florian; Dormann, Carsten F.. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography. https://doi.org/10.1111/ecog.02881

13. 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

14. Xiao, Xin-Yu; Tian, Ye; Yin, Erwei; He, Zhixian; Wang, Shiqi; Liu, Yalei; Xia, Qianchen. (2026). Lunar-Bench: Towards Evaluating Task-Oriented Reasoning of LLMs in Lunar Exploration Scenarios. Findings of the Association for Computational Linguistics: ACL 2026. https://aclanthology.org/2026.findings-acl.83/.

15. Nagihara, Seiichi; Nakamura, Yosio. (2018a). Apollo 15 ALSEP ARCSAV Heat Flow Experiment Calibrated Gradient Bridge Temperatures Bundle,version1.0. NASA Planetary Data System. Edited by D.R.Williams and S.A.McLaughlin. https://doi.org/10.17189/1518442

16. Diebold, Francis X.; Mariano, Roberto S. (1995). Comparing Predictive Accuracy. Journal of Business & Economic Statistics. https://doi.org/10.1080/07350015.1995.10524599

17. Gebru, Timnit; Morgenstern, Jamie; Vecchione, Briana; Vaughan, Jennifer Wortman; Wallach, Hanna; Daumé III, Hal; Crawford, Kate. (2021). Datasheets for datasets. Communications of the ACM. https://doi.org/10.1145/3458723

18. Hyndman, Rob J.; Koehler, Anne B.. (2006). Another look at measures of forecast accuracy. International Journal of Forecasting. https://doi.org/10.1016/j.ijforecast.2006.03.001

19. Bergmeir, Christoph; Benítez, José M.. (2012). On the use of cross-validation for time series predictor evaluation. Information Sciences. https://doi.org/10.1016/j.ins.2011.12.028

20. St. Clair, Michael; Million, Chase; Siegler, Matthew. (2020). Improvements to the Apollo 15 and 17 heat flow experiment data archive. Planetary and Space Science. https://doi.org/10.1016/j.pss.2020.104944

21. Montero-Manso, Pablo; Hyndman, Rob J.. (2021). Principles and algorithms for forecasting groups of time series: Locality and globality. International Journal of Forecasting. https://doi.org/10.1016/j.ijforecast.2021.03.004

22. Andrews, Donald W. K.. (1991). Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation. Econometrica. https://doi.org/10.2307/2938229

23. Salinas, David; Flunkert, Valentin; Gasthaus, Jan; Januschowski, Tim. (2020). DeepAR: Probabilistic forecasting with autoregressive recurrent networks. International Journal of Forecasting. https://doi.org/10.1016/j.ijforecast.2019.07.001

24. Clark, Todd E; McCracken, Michael W. (2001). Tests of equal forecast accuracy and encompassing for nested models. Journal of Econometrics. https://doi.org/10.1016/s0304-4076(01)00071-9

25. Ben Taieb, Souhaib; Bontempi, Gianluca; Atiya, Amir F.; Sorjamaa, Antti. (2012). A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition. Expert Systems with Applications. https://doi.org/10.1016/j.eswa.2012.01.039

26. Hunter, John D.. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering. https://doi.org/10.1109/mcse.2007.55

27. Newey, Whitney K.; West, Kenneth D. (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. Econometrica. https://doi.org/10.2307/1913610

28. Solon, Gary; Haider, Steven J.; Wooldridge, Jeffrey M.. (2015). What Are We Weighting For?. Journal of Human Resources. https://doi.org/10.3368/jhr.50.2.301.

29. St. Clair, Michael; Million, Chase; Siegler, Matthew. (2019). Apollo 15 and17 Heat Flow Experiment Concatenated Data Sets Bundle,version1.0; data collectionversion2.0. Edited by Jennifer Ward. NASA Planetary Data System. https://doi.org/10.17189/1518441

30. Kassis, Timothy; Agarwal, Vinayak; He, Yuhuan; Patel, Darshil; Brueckner, Aubrey M.. (2026). Scientific Agent Skills: A Library of Procedural Knowledge for Research Agents. arXiv preprint. https://arxiv.org/abs/2609.00065