Supplementary Material

From Faulty Greenhouse Sensors to Usable Training Data: An End-to-End Sensor Health Management Pipeline for Fault Detection, Data Repair, and Downstream Validation

M. A. Lamhour, M. Msalek, M. Kasbouya, S. Ardchir, M. Azzouazi

Tables and figures numbered with the prefix S belong to this document; all other table, figure and section numbers refer to the main paper. Every number reported here is produced by the released code from the released data.

S1 Detection benchmark (RQ2)

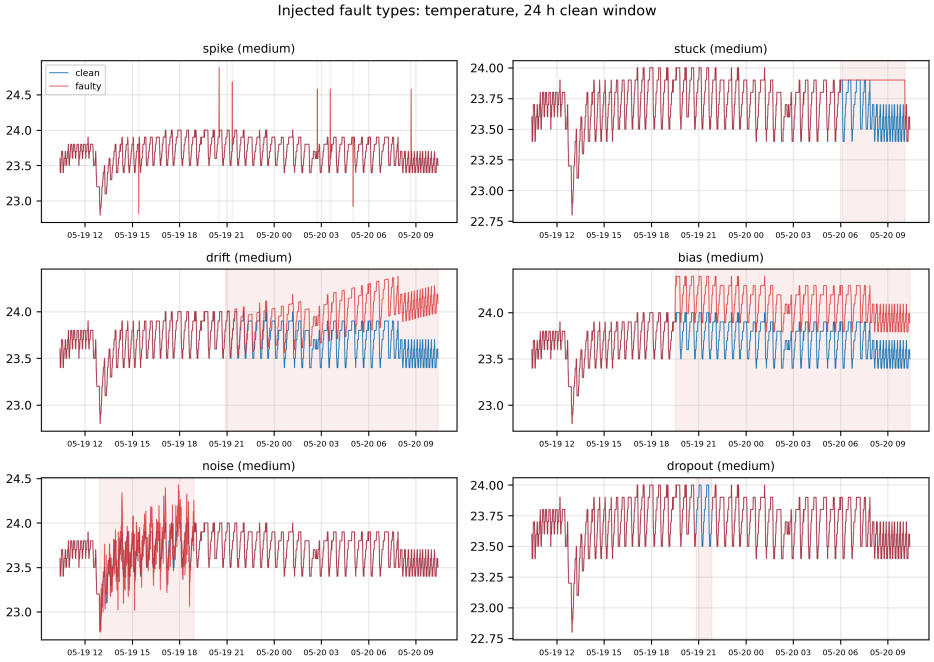


Figure S1. Injected fault types (medium severity) on a 24-h clean air-temperature window; the vertical axis of every panel is air temperature (°C). Shaded regions mark the ground-truth fault labels.

Threshold sensitivity

Table S1 repeats the benchmark of Table 8 for every threshold setting. Rules with the detector's Hampel constant $k_H \in \{4, 6, 8\}$, evaluated on the fixed default clean windows, give macro-F1 0.45, 0.44 and 0.43 and LSTM-AE quantiles 99–99.9% give 0.37, 0.36 and 0.35, with false-positive rates on the non-injected clean windows below 0.5% on average at every setting (worst window 1.11%). Isolation Forest rises from 0.29 at 0.5% contamination to 0.41 at 5%, above the default autoencoder over all six classes and above the default rules over the five value-corrupting classes (0.40 vs. 0.33), but its clean-window false-positive rate then reaches 11.6% on average and 18.8% on one window.

Window lengths

Window lengths matter less than thresholds (Table S2). Varying the Hampel window changes the rules' macro-F1 between 0.42 and 0.44; the autoencoder sequence length (each retrained) gives 0.38, 0.37 and 0.28, the drop at 120 min coming from spikes; the Isolation Forest rolling window gives 0.38, 0.33 and 0.31. All clean-window false-positive rates are at most 1% on average. The rules remain the best detector at every window length, while the autoencoder and Isolation Forest trade places at the shorter windows.

Table S1. Threshold sensitivity of the detectors on the 72 held-out scenarios. Macro-F1 over all six classes and over the five value-corrupting classes, per-class F1, and the false-positive rate (fraction of flagged samples) on the non-injected held-out clean windows (mean and worst window). Default operating points in bold.

Detector	Setting	Macro-F1	Macro-F1 (5)	bias	drift	noise	spike	stuck	FP mean	FP max
Rules	$k_H = 4$	0.450	0.344	0.002	0.090	0.017	0.628	0.985	0.09%	0.28%
Rules	$\mathbf{k_H = 6}$	0.444	0.333	0.000	0.073	0.009	0.591	0.991	0.00%	0.00%
Rules	$k_H = 8$	0.426	0.312	0.000	0.063	0.008	0.497	0.994	0.00%	0.00%
Isolation Forest	contamination 0.5%	0.286	0.145	0.011	0.000	0.163	0.373	0.179	0.07%	0.14%
Isolation Forest	**contamination 1%**	0.333	0.204	0.037	0.006	0.228	0.284	0.467	0.12%	0.28%
Isolation Forest	contamination 2%	0.376	0.295	0.162	0.053	0.478	0.171	0.613	2.26%	4.31%
Isolation Forest	contamination 5%	0.407	0.395	0.389	0.192	0.614	0.059	0.721	11.55%	18.82%
LSTM-AE	quantile 99%	0.369	0.253	0.219	0.196	0.229	0.596	0.025	0.33%	1.11%
LSTM-AE	**quantile 99.5%**	0.356	0.232	0.159	0.170	0.194	0.618	0.020	0.14%	0.42%
LSTM-AE	quantile 99.9%	0.354	0.227	0.132	0.144	0.161	0.681	0.016	0.05%	0.14%

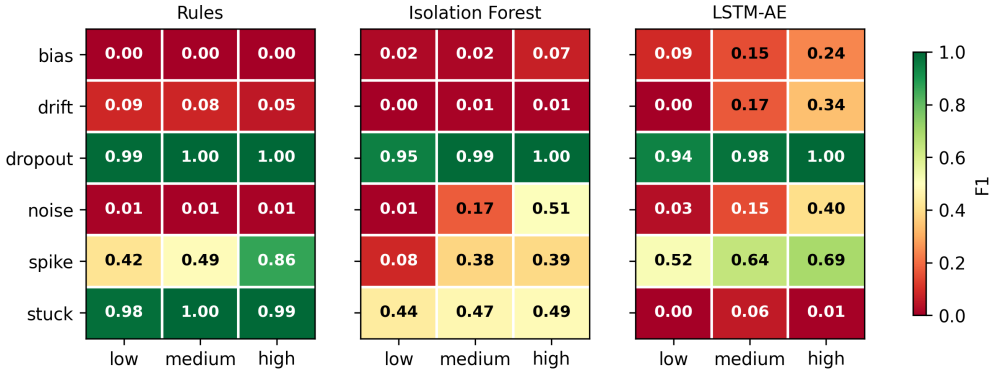


Figure S2. F1 by fault severity for each detector at the default operating points (mean over held-out scenarios).

Table S2. Window-length sensitivity of the detectors on the 72 held-out scenarios (other settings at their defaults). Columns as in Table S1; default lengths in bold. The LSTM-AE row at 60 min comes from the retrained model of this analysis and differs slightly from the default model of Table 8.

Detector	Window	Macro-F1	Macro-F1 (5)	bias	drift	noise	spike	stuck	FP mean	FP max
Rules (Hampel window)	31 min	0.421	0.310	0.001	0.092	0.007	0.458	0.991	0.10%	0.35%
Rules (Hampel window)	**61 min**	0.444	0.333	0.000	0.073	0.009	0.591	0.991	0.00%	0.00%
Rules (Hampel window)	121 min	0.422	0.314	0.006	0.041	0.016	0.519	0.986	0.21%	0.83%
Isolation Forest (rolling window)	15 min	0.381	0.264	0.064	0.025	0.316	0.442	0.470	0.19%	0.62%
Isolation Forest (rolling window)	**30 min**	0.333	0.204	0.037	0.006	0.228	0.284	0.467	0.12%	0.28%
Isolation Forest (rolling window)	60 min	0.307	0.180	0.034	0.001	0.199	0.261	0.406	0.69%	2.78%
LSTM-AE (sequence length, retrained)	30 min	0.384	0.270	0.199	0.188	0.270	0.675	0.016	0.24%	0.69%
LSTM-AE (sequence length, retrained)	**60 min**	0.372	0.249	0.156	0.170	0.203	0.692	0.022	0.07%	0.14%
LSTM-AE (sequence length, retrained)	120 min	0.275	0.151	0.133	0.125	0.126	0.357	0.012	0.99%	2.57%

Clean screen

The false-positive rate of the rules on the clean windows is exactly zero at $k_H = 6$ because the clean screen excludes the Hampel-flagged neighbourhoods of the same rule family. The rate at the stricter detector setting $k_H = 4$ (0.09%) shows that few additional rule alarms occur inside these windows, but it does not bound the selection bias itself, which arises from which windows the screen admits. Changing the screen's own k_H changes the window set: the screen yields no humidity 24-h window at $k_H = 4$, five at the default $k_H = 6$ (four are used) and twelve at $k_H = 8$, while the temperature and soil-moisture window counts are unchanged. The benchmark therefore cannot be repeated with the $k_H = 4$ screen, and Table S1 varies only the detector's k_H. A third sensitivity concerns the constancy criterion. With exact constancy evaluated on the raw samples (longer than 1 h, split at gaps longer than 5 min) instead of on 1-min medians, the share of clean soil-moisture minutes rises from 24.4% to 48.1% and the number of soil-moisture 24-h windows from three to nine, six of which precede the RQ4 boundary (regime R2, from 29 April); the humidity windows (five) and the temperature windows (two, none before the boundary) are

unchanged. The benchmark keeps the median-based criterion as the conservative choice: it rejects many healthy minutes of the slow index, but none of the windows it selects contains a minute that is stuck on the raw samples. Varying the spike exclusion margin (15, 30 or 60 min) leaves the window counts unchanged and keeps every detector's false-positive rate on the held-out clean windows below 0.5% (rules $\leq$0.07%). To measure the rules' flag rate on windows that were *not* selected by the rule family, 24-h windows were also screened with the autoencoder alone (present, in range, and less than 1% of slots flagged by the LSTM-AE, no stuck or Hampel criterion). On the eight humidity windows so selected the rules flag 0.70% of slots (Hampel spikes), and on the twelve soil-moisture windows 6.5%; the latter are mostly the real in-range stuck-at episodes of 5 and 7 May (5.5 and 4.3 h), which the autoencoder does not see (stuck-at F1 0.02). Evaluated with the constant-run rule on 1-min medians, the same windows would be flagged at 63.8%, the resolution artifact of Section 3.2.

Near-threshold and quasi-stuck scenarios

Table S3. Near-threshold stuck and quasi-stuck scenarios on the four held-out windows (three seeded repetitions each; default operating points). P = point-wise precision, F1 = point-wise F1, Ev = share of episodes with at least one flagged sample.

	Rules			Isolation Forest			LSTM-AE		
Scenario	P	F1	Ev	P	F1	Ev	P	F1	Ev
Stuck 30 min	0.00	0.00	0.00	0.45	0.23	0.50	0.00	0.00	0.00
Stuck 60 min	0.25	0.25	0.25	0.48	0.37	0.50	0.00	0.00	0.00
Stuck 90 min	0.98	0.99	1.00	0.47	0.42	0.50	0.00	0.00	0.00
Quasi-stuck 4 h	0.08	0.00	0.08	0.05	0.00	0.08	0.00	0.00	0.00
Quasi-stuck 8 h	0.04	0.00	0.08	0.16	0.03	0.17	0.06	0.01	0.08

S2 Repair benchmark (RQ3)

Table S4 gives the native-unit errors of all seven methods behind the equal-sensor nRMSE of Table 9. Standardizing the channels before KNN or iterative imputation changes nRMSE by less than 0.05 at every gap length and does not alter any ranking.

KNN distance ties. Because the channels are quantized, the 10th and 11th nearest donors are equidistant for 83% of single-sensor gap rows, so the donors KNN selects depend on the tie-breaking order of the neighbour search, and individual KNN errors can differ between software environments. Breaking the ties at random (a jitter of 10^{-6} standard deviations, five seeds) changed the equal-sensor KNN nRMSE by at most 0.025 and no ranking in our software environment; this is an empirical result for that environment, not a guaranteed bound. The released code pins the library versions.

Policy selection on pre-boundary soil-moisture windows. With the raw-sample constancy criterion, six soil-moisture windows precede the RQ4 boundary (Section S1). Repeating the benchmark of Section 4.3 on them (the two earliest build the diurnal profile and the other four are held out; same gap lengths, regimes, repeats and gap seed, normalization by the standard deviation of the held-out clean signal, 2.50 index units, and the selection rule of Table 10) selects linear interpolation in every regime and gap length. This agrees with Table 10 except for 12-h single-sensor gaps, where Table 10 selects KNN on the humidity windows and these soil-moisture windows select linear interpolation, with KNN not significantly different from it. Tables 9 and 10 are unchanged.

S3 Downstream evaluation (RQ4)

Error decomposition by training support

Figure S4 shows the decomposition reported numerically in Table 11. The split is post hoc: it uses the true value of the target and explains where the error of the original level-predicting LSTMs comes from, but it is not an operational rule. Because the target is unknown at prediction time, Table S9 repeats the decomposition with the test samples grouped by the last observed value $y(t)$, which is known when the forecast is made. The repair-versus-raw differences are the same under both groupings (about -4 index units at every horizon, intervals excluding zero). Reaching the persistence level is not: grouped by the last observed value, only rules-guided repair stays at that level at 5 min ($+0.21$, interval -0.02 to 0.43), and at 1 h and 6 h every repaired variant is worse than persistence. The two groupings disagree only where the target lies at 48–53, around the support edge of 50 where the level models saturate near the top of their training range.

Table S4. Per-sensor imputation RMSE in native units (mean over repeats and held-out windows) for all seven methods; bold = best per sensor, regime and gap length. Humidity in % RH, temperature in °C, soil-moisture index in index units. The hybrid equals linear interpolation up to 1 h; in the node-outage regime the KNN and iterative imputers reduce to the column mean and coincide.

		Single-sensor gap					Node outage				
Sensor	**Method**	5 min	15 min	1 h	4 h	12 h	5 min	15 min	1 h	4 h	12 h
Humidity	forward-fill	0.14	0.52	0.81	1.59	4.15	0.30	**0.11**	0.54	6.36	4.81
	linear	**0.11**	**0.41**	**0.22**	**0.55**	3.74	**0.27**	**0.11**	**0.20**	**3.49**	**3.14**
	spline	0.36	0.90	0.68	0.82	11.97	0.35	0.25	0.37	4.20	3.55
	KNN	0.23	1.13	0.91	1.30	**1.94**	6.02	4.33	3.11	4.91	5.13
	iterative	1.73	2.20	2.24	2.87	4.51	6.02	4.33	3.11	4.91	5.13
	profile	2.60	1.75	2.53	3.01	4.43	2.97	1.96	2.54	4.55	3.71
	hybrid	**0.11**	**0.41**	**0.22**	3.01	4.43	**0.27**	**0.11**	**0.20**	4.55	3.71
Temperature	forward-fill	0.11	0.21	0.24	0.21	0.22	0.12	0.23	0.24	0.23	0.22
	linear	**0.06**	**0.15**	0.20	0.21	0.18	**0.05**	**0.18**	0.20	0.20	0.21
	spline	0.13	**0.15**	**0.16**	**0.16**	**0.17**	0.12	**0.18**	**0.16**	**0.15**	**0.17**
	KNN	0.13	0.16	0.15	0.17	0.18	0.16	0.21	**0.16**	0.18	0.20
	iterative	0.14	0.18	**0.16**	**0.16**	0.18	0.16	0.21	**0.16**	0.18	0.20
	profile	0.17	0.18	0.23	0.20	0.20	0.22	0.21	0.21	0.21	0.20
	hybrid	**0.06**	**0.15**	0.20	0.20	0.20	**0.05**	**0.18**	0.20	0.21	0.20
Soil-moisture index	forward-fill	1.79	1.66	1.63	1.67	1.60	1.71	1.79	1.53	1.69	1.82
	linear	1.47	1.57	**1.34**	1.44	1.44	1.37	1.59	1.36	1.42	**1.22**
	spline	**1.20**	1.53	1.50	5.08	5.96	**1.21**	**1.37**	2.13	5.54	14.30
	KNN	**1.20**	**1.24**	**1.34**	1.31	**1.27**	1.39	1.44	**1.21**	**1.24**	1.33
	iterative	1.21	1.27	4.32	2.32	1.99	1.39	1.44	**1.21**	**1.24**	1.33
	profile	1.46	1.73	1.63	1.48	1.46	1.50	1.85	1.33	1.49	1.64
	hybrid	1.47	1.57	**1.34**	1.48	1.46	1.37	1.59	1.36	1.49	1.64

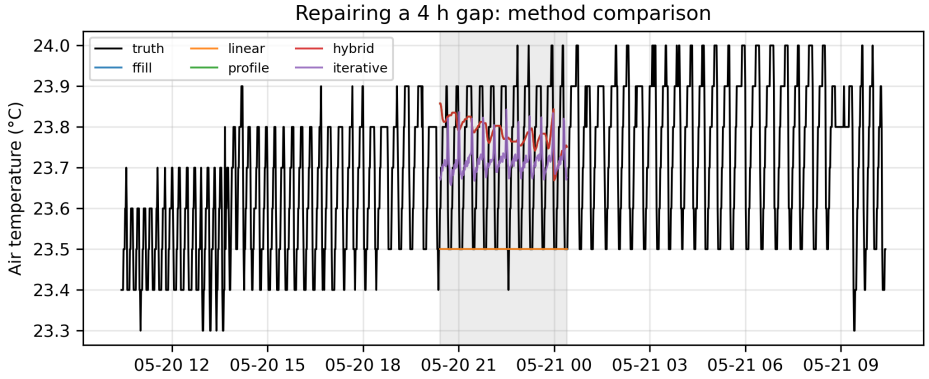


Figure S3. Qualitative comparison of imputation methods repairing a 4-h gap in air temperature.

Rolling-origin blocks and the operational test set

Table S5 repeats the 5-min experiment of the original level LSTMs on three chronological blocks, each with a screened and an operational test set (in the operational set, missing input slots carry the last earlier observation forward, never a later one). Two results hold in all six block-and-set combinations: on the full test sets persistence is the best forecast (screened RMSE 0.23, 1.24 and 1.29; operational 0.22, 0.92 and 3.21), and inside the training support rules-guided repair beats raw training (screened 0.70, 1.16, 1.05 vs. 1.48, 2.55, 4.91; operational 1.27, 0.94, 1.60 vs. 3.27, 2.58, 6.37). The other comparisons are not uniform. On the full sets rules-guided repair beats raw training in five of six cases, the exception being the operational set of the third block (13.55 vs. 13.37); autoencoder-guided repair beats raw training in four of six and is worse in both sets of the third block. In the second and third blocks deletion is worse than every repaired variant (screened 3.33 vs. 2.62–2.80 and 13.48 vs. 9.90–11.19); in the second block it is still better than raw training (3.33 vs. 4.21), in the third it is worse (13.48 vs. 10.94). In the first block, which follows a mostly healthy training period, it is close to the best learned variant (screened 0.78, within 0.01 of rules-guided repair; operational 1.17 vs. 1.25). Inside the support the repaired models are close to persistence in the second block (screened 1.16–1.18 vs. 1.10), at or below it in the third (1.05–1.30 vs. 1.34), and well above it in the quiet first block (0.70–0.82 vs. 0.21). Across the three screened blocks rules-guided repair beats raw training in 3 of 3 (mean difference -1.12, paired t $p = 0.03$) and autoencoder-guided repair in 2 of 3 (-0.77, $p = 0.25$). In the third block every error is larger on the operational set than on the screened set, persistence included (3.21 vs. 1.29). In the first block, operational RMSE increases for

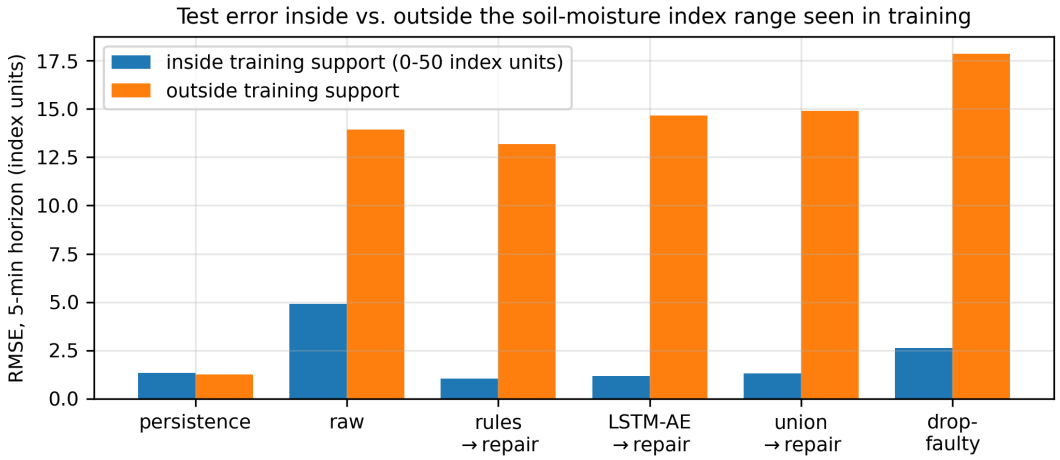


Figure S4. Test RMSE of the original level-predicting LSTMs at the 5-min horizon, decomposed by whether the target lies inside the range of screened-clean training targets (0–50 index units) or outside it (10 seeds, mean). Persistence is computed on the same samples.

Table S5. Rolling-origin evaluation of the original level-predicting LSTM (5-min horizon, 10 seeds). For each block every fitted component is re-estimated on the data preceding the block. "Screened" = the common untouched test set of Section 3.6 restricted to the block; "operational" = every observed target of the block, inputs taken as recorded with missing slots forward-filled inside the block and flagged inputs kept. Inside = RMSE on targets inside the training support of that block; share = fraction of targets inside the support.

		Screened test set				Operational test set			
Block	**Training data**	n	RMSE	SS	inside (share)	n	RMSE	SS	inside (share)
B1: 5–8 May (train 18 Apr–5 May)	Persistence	343	0.23	0.00	0.21 (86%)	864	0.22	0.00	0.21 (83%)
	Raw	343	1.61 ± 0.63	-6.13	1.48 (86%)	864	3.14 ± 2.56	-13.24	3.27 (83%)
	Rules→repair	343	0.78 ± 0.23	-2.45	0.70 (86%)	864	1.25 ± 0.21	-4.65	1.27 (83%)
	LSTM-AE→repair	343	0.83 ± 0.23	-2.70	0.76 (86%)	864	2.28 ± 0.76	-9.34	2.44 (83%)
	Union→repair	343	0.90 ± 0.23	-2.99	0.82 (86%)	864	2.25 ± 0.71	-9.23	2.40 (83%)
	Drop-faulty	343	0.78 ± 0.35	-2.47	0.63 (86%)	864	1.17 ± 0.33	-4.31	1.11 (83%)
B2: 8–11 May (train 18 Apr–8 May)	Persistence	443	1.24	0.00	1.10 (43%)	844	0.92	0.00	0.76 (49%)
	Raw	443	4.21 ± 1.20	-2.38	2.55 (43%)	844	3.78 ± 1.02	-3.14	2.58 (49%)
	Rules→repair	443	2.72 ± 0.55	-1.19	1.16 (43%)	844	2.56 ± 0.54	-1.80	0.94 (49%)
	LSTM-AE→repair	443	2.62 ± 0.42	-1.11	1.16 (43%)	844	2.50 ± 0.44	-1.73	0.99 (49%)
	Union→repair	443	2.80 ± 0.57	-1.25	1.18 (43%)	844	2.63 ± 0.58	-1.87	0.96 (49%)
	Drop-faulty	443	3.33 ± 0.38	-1.68	1.20 (43%)	844	3.16 ± 0.37	-2.45	1.18 (49%)
B3: 12–22 May (train 18 Apr–12 May)	Persistence	1162	1.29	0.00	1.34 (44%)	1565	3.21	0.00	3.70 (35%)
	Raw	1162	10.94 ± 3.45	-7.48	4.91 (44%)	1565	13.37 ± 3.77	-3.17	6.37 (35%)
	Rules→repair	1162	9.90 ± 0.88	-6.67	1.05 (44%)	1565	13.55 ± 1.12	-3.22	1.60 (35%)
	LSTM-AE→repair	1162	11.01 ± 1.00	-7.53	1.17 (44%)	1565	15.03 ± 1.27	-3.69	1.56 (35%)
	Union→repair	1162	11.19 ± 1.11	-7.67	1.30 (44%)	1565	15.24 ± 1.44	-3.75	1.66 (35%)
	Drop-faulty	1162	13.48 ± 0.97	-9.45	2.62 (44%)	1565	17.85 ± 1.26	-4.57	2.78 (35%)

every learned variant, while persistence changes little (0.22 vs. 0.23); in the second block, operational RMSE is lower for every method. Inside the support of the third block the repaired models have lower RMSE than persistence on the operational set (1.56–1.66 vs. 3.70), because persistence inherits the faulty targets while the repaired models partly smooth over them. With three blocks these across-block statistics have little power and serve as a consistency check.

Random-mask control

Design. The protocol was fixed before any of its models was trained and is released with the code. On the training partition only, and separately for each input channel, the observed values flagged by the rules (the rules mask minus the genuine missing slots, which stay masked identically in both conditions) are replaced by randomly placed masks with the same number of values and the same multiset of contiguous run lengths. Runs are placed uniformly among feasible positions on observed training slots, without overlap and with a one-slot separation, and may cover naturally faulty values by chance. Repair, normalization, model, training and the common test set are those of rules→repair. Five mask seeds are crossed with the ten training seeds (50 fits)

Table S6. Random-mask control for the original level-predicting LSTM (5-min horizon, common screened test set). Random-mask→repair masks, on the training partition only, the same number of observed values per channel as the rule detector, in runs of the same lengths, at random positions (5 mask seeds × 10 training seeds); genuine missing values and every other step are identical to rules→repair. In the split variant the single 2,026-slot soil-moisture run is placed as two 1,013-slot runs. RMSE: mean ± sd over fits. Δ = pooled RMSE of the control minus that of rules→repair (positive favors rules) with the 95% bootstrap interval over contiguous runs of test samples; "masks" = number of the five mask seeds whose seed-averaged RMSE is above that of rules→repair. OOR left = share of the out-of-range soil-moisture episode the control leaves unmasked in training.

Training data	RMSE full	RMSE inside	RMSE outside	Δ full [95% CI]	Δ inside [95% CI]	Masks full / inside	OOR left
Persistence (reference)	1.29	1.34	1.25	n/a	n/a	n/a	n/a
Raw (reference)	10.94 ± 3.45	4.91 ± 1.79	13.93 ± 4.36	n/a	n/a	n/a	100%
Rules→repair	9.90 ± 0.88	1.05 ± 0.12	13.19 ± 1.17	n/a	n/a	n/a	0%
Random-mask→repair	7.36 ± 1.94	1.29 ± 0.73	9.74 ± 2.57	-2.34 [-5.11, -0.62]	+0.42 [+0.31, +0.72]	0 / 3	7–10%
Random-mask→repair, split	6.27 ± 2.30	2.05 ± 0.96	8.16 ± 2.99	-3.27 [-8.43, -0.13]	+1.20 [+0.96, +1.75]	0 / 5	48–53%

at the 5-min horizon, and the difference to rules→repair is analysed across mask seeds, across training seeds and, for temporal sampling, with the run-based bootstrap on pooled losses. Count and run lengths matched the rules mask exactly for every channel and seed, no fit failed, and rules→repair and raw reproduced Table 11.

Structural limitation and split control. The longest soil-moisture run of the rules mask (2,026 slots, the out-of-range episode) fits only in the observed stretch that contains it, so every random placement of the primary control re-masks 90–93% of that episode. A secondary control, also fixed before training, places this run as two 1,013-slot runs; it still re-masks part of the episode (it leaves 48–53% of it unmasked), so neither control fully separates the targeting of the dominant fault from the amount of masking.

Results. On the prespecified primary outcome, the full screened test set, the random control has lower error than rules-guided repair for all five mask seeds (pooled difference -2.34, interval -5.11 to -0.62), and so does the split control (Table S6). The advantage comes from the samples outside the training support. Models trained on rules-repaired data saturate near the top of the training range (slope of their predictions on the last observed value 0.16 outside the support), whereas the random masks leave 7–10% (split: 48–53%) of the out-of-range episode in the training data, and the models trained on them follow their lagged input more closely (slopes 0.40 and 0.60). This pattern is consistent with the large steps of that episode helping extrapolation to the unseen post-outage level, although the control was not designed to establish that mechanism. Inside the training support the comparison reverses: rules-guided repair has lower error than the split control for all five mask seeds and all ten training seeds when averaged over the other factor (44 of 50 individual fits; pooled difference $+1.20$, interval 0.96 to 1.75), while against the primary control the result depends on the mask realization (three of five masks favor rules; pooled $+0.42$). The control therefore does not show a benefit of targeting on the full test set; inside the support the results are consistent with masking the dominant soil-moisture fault mattering more than masking the same amount of data elsewhere.

Residual LSTM

The residual LSTM is the LSTM of Table 11 trained on the change $y(t+H) - y(t)$ with the absolute level still among its inputs. It reduces the 5-min error of the level models from 9.9–13.5 to 1.3–1.8 but does not beat persistence on the full test set (Table S7). It was evaluated before the models of Table 14.

Table S7. Residual LSTM (LSTM of Table 11 trained on the change $y(t+H) - y(t)$, prediction = last observed value + predicted change, absolute level still among the inputs). RMSE mean ± sd over 10 seeds on the common test set.

	RMSE full			RMSE inside support		
Training data	5 min	1 h	6 h	5 min	1 h	6 h
Persistence	1.29	1.53	1.97	1.34	1.52	1.58
Raw	1.30 ± 0.00	1.70 ± 0.19	2.55 ± 0.55	1.31	1.47	2.35
Rules→repair	1.81 ± 0.44	1.66 ± 0.09	2.08 ± 0.15	1.18	1.48	1.34
LSTM-AE→repair	1.79 ± 0.74	1.65 ± 0.08	2.44 ± 0.35	1.25	1.50	1.57
Union→repair	1.82 ± 0.69	1.67 ± 0.09	2.59 ± 0.31	1.23	1.46	1.66
Drop-faulty	1.30 ± 0.01	1.55 ± 0.02	2.03 ± 0.05	1.32	1.49	1.53

Training seeds

Paired over the ten seeds, rules-guided repair has lower RMSE than autoencoder-guided repair in 10 of 10 seeds (exact Wilcoxon $p = 0.002$) and autoencoder-guided repair beats drop-faulty in 10 of 10 ($p = 0.002$), while the autoencoder-versus-raw comparison is a coin toss on the full set (6 of 10, $p = 0.85$) because the raw model's seed-to-seed variation (sd 3.5) dwarfs the differences. These seed comparisons quantify training stochasticity and are not used for inference about forecast accuracy.

Sensitivity of the temporal intervals

Every interval of Table 12 was repeated with block lengths of 12 and 48 samples and with a bootstrap over whole contiguous runs of retained samples, so that no block joins two non-adjacent samples (Table S8); the block-24 column of Table S8 is the interval of Table 12 itself (same bootstrap draw). Inside the training support, which contains about 100 such runs, every conclusion of Section 4.4 holds under all four schemes, with two borderline limits at the longest block length: autoencoder-guided repair against persistence at 5 min (upper limit $+0.005$) and union-guided repair against persistence at 1 h (lower limit -0.005). The deletion-versus-raw difference at 6 h, not significant in Table 12, is significant only with the shortest blocks. On the full test set the moving-block conclusions also hold, with one exception of the same kind: the deletion-versus-raw difference at 5 min, whose interval in Table 12 includes zero (-0.11 to $+3.35$), excludes zero with 12-sample blocks ($+0.66$ to $+3.06$). The run-based bootstrap rests on only 12–14 runs and its intervals for the repaired-versus-raw differences include zero; full-set comparisons between training variants of the level models are therefore indicative, and the inside-support comparisons are the primary evidence for them.

Table S8. Sensitivity of the temporal intervals of Table 12 to the resampling scheme: pooled RMSE difference (first minus second) with 95% moving-block bootstrap intervals for block lengths of 12, 24 and 48 samples, and with a bootstrap over contiguous runs of retained test samples (n runs). The difference and the block-24 column are those of Table 12 (same bootstrap draw, which also gives the union rows not shown there); the block-12, block-48 and run-based columns are separate draws. Limits within 0.01 of zero are shown with three decimals.

H	Set	Comparison	Δ	block 12	block 24	block 48	runs	n
5 min	full	raw vs. persistence	+10.13	[+8.80, +11.34]	[+8.17, +11.78]	[+7.46, +12.31]	[+6.47, +18.46]	12
		rules→repair vs. persistence	+8.65	[+6.43, +10.23]	[+5.56, +10.49]	[+4.35, +11.06]	[+2.74, +20.37]	12
		LSTM-AE→repair vs. persistence	+9.76	[+7.31, +11.54]	[+6.25, +11.95]	[+4.70, +12.52]	[+3.36, +22.32]	12
		union→repair vs. persistence	+9.95	[+7.60, +11.73]	[+6.64, +12.12]	[+4.92, +12.64]	[+3.38, +22.47]	12
		drop-faulty vs. persistence	+12.22	[+9.69, +14.27]	[+8.44, +14.64]	[+6.89, +15.44]	[+4.92, +26.04]	12
		rules→repair vs. raw	-1.48	[-2.47, -0.73]	[-2.88, -0.75]	[-3.59, -0.74]	[-4.05, +2.42]	12
		LSTM-AE→repair vs. raw	-0.37	[-1.65, +0.41]	[-2.12, +0.54]	[-3.02, +0.57]	[-3.72, +4.54]	12
		union→repair vs. raw	-0.18	[-1.34, +0.67]	[-1.96, +0.81]	[-2.83, +0.84]	[-3.53, +4.65]	12
		drop-faulty vs. raw	+2.10	[+0.66, +3.06]	[-0.11, +3.35]	[-0.97, +3.39]	[-1.72, +8.01]	12
		LSTM-AE→repair vs. rules→repair	+1.11	[+0.87, +1.29]	[+0.75, +1.32]	[+0.62, +1.38]	[+0.46, +2.31]	12
5 min	inside	raw vs. persistence	+3.85	[+2.91, +4.70]	[+2.49, +4.92]	[+1.99, +5.30]	[+2.71, +6.01]	106
		rules→repair vs. persistence	-0.28	[-0.41, -0.15]	[-0.43, -0.14]	[-0.47, -0.13]	[-0.40, -0.18]	106
		LSTM-AE→repair vs. persistence	-0.16	[-0.29, -0.03]	[-0.31, -0.01]	[-0.35, +0.005]	[-0.27, -0.02]	106
		union→repair vs. persistence	-0.02	[-0.17, +0.14]	[-0.19, +0.15]	[-0.24, +0.19]	[-0.14, +0.17]	106
		drop-faulty vs. persistence	+1.30	[+1.00, +1.63]	[+0.85, +1.70]	[+0.72, +1.83]	[+0.89, +2.07]	106
		rules→repair vs. raw	-4.13	[-4.92, -3.21]	[-5.12, -2.85]	[-5.46, -2.29]	[-6.33, -3.06]	106
		LSTM-AE→repair vs. raw	-4.01	[-4.76, -3.06]	[-5.04, -2.71]	[-5.42, -2.24]	[-6.11, -2.90]	106
		union→repair vs. raw	-3.87	[-4.63, -3.00]	[-4.87, -2.63]	[-5.12, -2.16]	[-5.91, -2.79]	106
		drop-faulty vs. raw	-2.55	[-3.10, -1.84]	[-3.26, -1.66]	[-3.52, -1.20]	[-3.95, -1.81]	106
		LSTM-AE→repair vs. rules→repair	+0.12	[+0.11, +0.14]	[+0.10, +0.15]	[+0.09, +0.15]	[+0.10, +0.18]	106
1 h	full	raw vs. persistence	+9.66	[+8.27, +10.83]	[+7.80, +11.22]	[+6.90, +11.74]	[+6.42, +17.44]	13
		rules→repair vs. persistence	+8.52	[+6.36, +10.12]	[+5.53, +10.46]	[+4.04, +10.80]	[+3.24, +19.66]	13
		LSTM-AE→repair vs. persistence	+8.99	[+6.76, +10.62]	[+5.62, +11.05]	[+4.21, +11.54]	[+3.35, +20.79]	13
		union→repair vs. persistence	+9.02	[+6.68, +10.65]	[+5.77, +11.18]	[+4.32, +11.59]	[+3.40, +20.55]	13
		drop-faulty vs. persistence	+10.69	[+8.03, +12.68]	[+7.00, +13.15]	[+5.00, +13.63]	[+4.18, +24.21]	13
		rules→repair vs. raw	-1.14	[-2.25, -0.40]	[-2.66, -0.38]	[-3.34, -0.37]	[-3.70, +2.82]	13
		LSTM-AE→repair vs. raw	-0.67	[-1.86, +0.16]	[-2.51, +0.17]	[-3.36, +0.27]	[-3.51, +3.87]	13
		union→repair vs. raw	-0.63	[-1.87, +0.17]	[-2.34, +0.21]	[-3.10, +0.27]	[-3.54, +3.91]	13
		drop-faulty vs. raw	+1.04	[-0.38, +2.06]	[-0.99, +2.17]	[-2.16, +2.29]	[-2.61, +6.83]	13
		LSTM-AE→repair vs. rules→repair	+0.46	[+0.32, +0.57]	[+0.24, +0.60]	[+0.10, +0.64]	[+0.02, +1.13]	13
1 h	inside	raw vs. persistence	+4.06	[+3.00, +4.88]	[+2.61, +5.24]	[+2.11, +5.56]	[+2.96, +6.17]	107
		rules→repair vs. persistence	+0.35	[+0.11, +0.50]	[+0.07, +0.57]	[+0.01, +0.61]	[+0.15, +0.74]	107
		LSTM-AE→repair vs. persistence	+0.15	[-0.06, +0.29]	[-0.07, +0.31]	[-0.12, +0.35]	[-0.002, +0.43]	107
		union→repair vs. persistence	+0.26	[+0.08, +0.42]	[+0.05, +0.46]	[-0.005, +0.51]	[+0.11, +0.58]	107
		drop-faulty vs. persistence	+0.59	[+0.32, +0.84]	[+0.23, +0.93]	[+0.12, +0.99]	[+0.32, +1.15]	107
		rules→repair vs. raw	-3.71	[-4.42, -2.86]	[-4.61, -2.55]	[-4.88, -2.08]	[-5.49, -2.69]	107

Table S8 (continued)

H	Set	Comparison	Δ	block 12	block 24	block 48	runs	n
		LSTM-AE→repair vs. raw	-3.91	[-4.64, -2.99]	[-4.88, -2.69]	[-5.13, -2.20]	[-5.79, -2.84]	107
		union→repair vs. raw	-3.79	[-4.47, -2.87]	[-4.73, -2.55]	[-4.98, -2.10]	[-5.69, -2.78]	107
		drop-faulty vs. raw	-3.47	[-4.04, -2.71]	[-4.24, -2.37]	[-4.50, -2.07]	[-5.10, -2.57]	107
		LSTM-AE→repair vs. rules→repair	-0.20	[-0.23, -0.15]	[-0.25, -0.13]	[-0.27, -0.11]	[-0.32, -0.14]	107
6 h	full	raw vs. persistence	+11.16	[+9.96, +12.65]	[+9.30, +13.30]	[+8.80, +13.84]	[+8.20, +18.05]	14
		rules→repair vs. persistence	+9.47	[+7.44, +11.24]	[+6.40, +11.73]	[+5.01, +12.64]	[+4.86, +19.36]	14
		LSTM-AE→repair vs. persistence	+9.53	[+7.42, +11.36]	[+6.43, +11.93]	[+4.89, +12.74]	[+4.74, +19.70]	14
		union→repair vs. persistence	+9.75	[+7.56, +11.54]	[+6.61, +12.20]	[+5.14, +13.11]	[+4.91, +20.22]	14
		drop-faulty vs. persistence	+12.49	[+10.22, +14.63]	[+9.34, +15.13]	[+7.79, +15.88]	[+7.59, +24.48]	14
		rules→repair vs. raw	-1.68	[-2.98, -0.74]	[-3.54, -0.69]	[-4.30, -0.67]	[-4.22, +2.41]	14
		LSTM-AE→repair vs. raw	-1.63	[-2.90, -0.64]	[-3.56, -0.46]	[-4.43, -0.56]	[-4.34, +2.63]	14
		union→repair vs. raw	-1.41	[-2.76, -0.42]	[-3.27, -0.35]	[-4.24, -0.33]	[-4.19, +2.71]	14
		drop-faulty vs. raw	+1.34	[-0.14, +2.41]	[-0.68, +2.64]	[-1.59, +2.67]	[-1.48, +6.36]	14
		LSTM-AE→repair vs. rules→repair	+0.05	[-0.02, +0.11]	[-0.06, +0.13]	[-0.14, +0.16]	[-0.18, +0.34]	14
6 h	inside	raw vs. persistence	+5.76	[+4.70, +6.78]	[+4.22, +7.17]	[+3.66, +7.70]	[+4.49, +8.24]	97
		rules→repair vs. persistence	+1.70	[+1.44, +1.99]	[+1.41, +2.08]	[+1.43, +2.18]	[+1.48, +2.17]	97
		LSTM-AE→repair vs. persistence	+1.38	[+1.12, +1.64]	[+1.10, +1.73]	[+1.14, +1.81]	[+1.18, +1.82]	97
		union→repair vs. persistence	+1.67	[+1.40, +1.94]	[+1.38, +2.04]	[+1.39, +2.17]	[+1.45, +2.15]	97
		drop-faulty vs. persistence	+4.67	[+4.48, +4.84]	[+4.48, +4.89]	[+4.58, +4.92]	[+4.56, +4.84]	97
		rules→repair vs. raw	-4.05	[-4.90, -3.17]	[-5.18, -2.77]	[-5.60, -2.13]	[-6.11, -2.98]	97
		LSTM-AE→repair vs. raw	-4.38	[-5.23, -3.45]	[-5.58, -3.09]	[-6.00, -2.52]	[-6.52, -3.30]	97
		union→repair vs. raw	-4.09	[-4.93, -3.14]	[-5.23, -2.84]	[-5.59, -2.27]	[-6.08, -3.02]	97
		drop-faulty vs. raw	-1.09	[-2.05, -0.10]	[-2.38, +0.45]	[-2.86, +0.97]	[-3.59, +0.10]	97
		LSTM-AE→repair vs. rules→repair	-0.33	[-0.35, -0.31]	[-0.36, -0.30]	[-0.36, -0.29]	[-0.37, -0.31]	97

Table S9. Inside-support errors of the original level-predicting LSTMs with the test samples grouped by the true target (post hoc, as in Table 12) or by the last observed value $y(t)$, which is known when the forecast is made; inside = value within the training support (0–50 index units). (a) Pooled RMSE in index units and the share of test samples on which the two groupings agree. (b) Pooled RMSE difference (first minus second) with the 95% moving-block bootstrap interval (block 24, 2,000 replicates, the procedure of Table 12; the target rows reproduce it). Values from script 24, section (j).

(a) RMSE inside the support

H	Grouping	Agree	n	Persistence	Raw	Rules	LSTM-AE	Union
5 min	target	82.1%	510	1.34	5.19	1.06	1.18	1.32
	last input		510	1.35	5.72	1.56	1.70	1.86
1 h	target	77.8%	510	1.52	5.58	1.87	1.67	1.78
	last input		498	1.47	6.08	2.44	2.24	2.36
6 h	target	72.7%	487	1.58	7.33	3.28	2.95	3.24
	last input		439	1.61	8.25	4.12	3.80	4.09

(b) Differences inside the support

H	Grouping	Rules − pers.	LSTM-AE − pers.	Union − pers.	Raw − pers.	Rules − raw	LSTM-AE − raw
5 min	target	-0.28 $[-0.43, -0.14]$	-0.16 $[-0.31, -0.01]$	-0.02 $[-0.19, +0.15]$	$+3.85$ $[+2.49, +4.92]$	-4.13 $[-5.12, -2.85]$	-4.01 $[-5.04, -2.71]$
	last input	$+0.21$ $[-0.02, +0.43]$	$+0.35$ $[+0.08, +0.60]$	$+0.51$ $[+0.22, +0.80]$	$+4.37$ $[+2.85, +5.54]$	-4.16 $[-5.09, -2.92]$	-4.01 $[-4.99, -2.76]$
1 h	target	$+0.35$ $[+0.07, +0.57]$	$+0.15$ $[-0.07, +0.31]$	$+0.26$ $[+0.05, +0.46]$	$+4.06$ $[+2.61, +5.24]$	-3.71 $[-4.61, -2.55]$	-3.91 $[-4.88, -2.69]$
	last input	$+0.97$ $[+0.52, +1.33]$	$+0.77$ $[+0.39, +1.10]$	$+0.90$ $[+0.51, +1.25]$	$+4.62$ $[+3.01, +5.85]$	-3.64 $[-4.50, -2.50]$	-3.84 $[-4.73, -2.63]$
6 h	target	$+1.70$ $[+1.41, +2.08]$	$+1.38$ $[+1.10, +1.73]$	$+1.67$ $[+1.38, +2.04]$	$+5.76$ $[+4.22, +7.17]$	-4.05 $[-5.18, -2.77]$	-4.38 $[-5.58, -3.09]$
	last input	$+2.51$ $[+2.07, +3.05]$	$+2.19$ $[+1.75, +2.72]$	$+2.48$ $[+2.03, +3.05]$	$+6.64$ $[+4.74, +8.13]$	-4.13 $[-5.18, -2.72]$	-4.45 $[-5.49, -3.00]$