Fault diagnosis of a water-cooled centrifugal chiller from a physics-structured Bayesian model without fault labels, on laboratory fault runs at five instrumentation levels under a pre-registered protocol

Ricardo Marquez
rmarquez123@gmail.com

November 2026

Abstract

Chiller fault diagnosis should name the fault, its severity and its confidence without labeled fault examples, which operators never have. A physics-structured Bayesian model of a water-cooled centrifugal chiller, fixed on normal operation alone, was run on the fault runs of a laboratory benchmark at five nested sensor sets, under a protocol registered before any fault run was read. This is an inference study: each run's readings are weighed by their evidence against eight hypotheses, and the answers are scored against a published rule-based comparator. Because the refrigerant-side model had failed its registered acceptance test, the registered fallback governs: the physics-based claims are confined to the two water-flow faults and the confirmatory comparison runs on 17 runs rather than 39. Two dated amendments after the normal-run stage dropped the classifier context arms for compute and repaired the comparator, which had stopped with an error on one published steady state. With full instrumentation the model names the registered class of 33 of 39 scored runs against 19 for the comparator, and still answers on five channels, where the comparator cannot run. On the registered confirmatory population the model's merged run-level accuracy is below the comparator's at every sensor set, and the difference is not distinguishable from zero at this sample size. Its stated confidence is calibrated at the two registered sets. The flow-fault magnitude intervals are narrow, but seven of nine sit below the nominal severities, so the registered coverage rule fails. No advance detection prediction was contradicted, and no verdict changes under a second linear-algebra kernel. Fault labels are thus not needed to name faults with a trustworthy confidence, but magnitudes need a measured reference before they are quoted. The scope is one laboratory machine with few runs per fault family; the cost, one run of the registered computation.

1 Introduction

Vapor-compression chillers are the largest single electrical load of most commercial buildings, and they degrade quietly: condenser tubes foul, refrigerant leaks past seals, oil migrates out of the compressor sump, air enters the low-pressure side, and a pump or a strainer throttles a water loop. None of these stops the machine; each raises its power draw for the same delivered cooling by amounts that accumulate over a season and that an operator has no direct instrument for. Automated fault detection and diagnosis reads those conditions from the sensors a chiller already carries, and the difficulty is that the readings move for innocent reasons — load and entering water temperature — far more than they move for a fault. Katipamula and Brambley (2005) survey the

field and its evaluation problem. The community's reference data set is the ASHRAE research project RP-1043 of Comstock and Braun (1999), in which a 90-ton water-cooled centrifugal chiller was run over a grid of loads and water temperatures, and then run again with seven faults introduced at graded severities, each against a matched normal run.

Methods evaluated on that data set fall into three groups. Rule-based methods compare each reading with a regression of normal operation and match the signs of the residuals against a table of expected symptoms; Reddy (2007) evaluated four such methods on these runs under a common protocol and ranked a five-feature regression method the best, while reporting that none of them resolved the mildest severities. Data-driven classifiers learn the fault label from the readings (Han et al. 2011); they are strong on the data they were trained on, need labeled fault examples, and are hard to audit. Model-based probabilistic methods write the physics down and compare competing explanations of the same readings; Bayesian-network diagnosis on these runs is established work (Zhou et al. 2009; Zhao, Xiao and Wang 2013), and Zhao et al. (2013) report a statistical method on them. Nothing in this report is a first of its kind. Two evaluation hazards shape what follows: a method tuned while its fault-run scores are visible cannot be evaluated on those scores, and a benchmark split by measured points rather than by whole runs lets a method recognize the run it is asked about.

TERRA is a general-purpose engine for probabilistic inference on physical networks. It treats a network as a graph on which the physics is written locally: a conservation law at every node, a constitutive law on every edge, and, attached to the same graph, whatever the sensors report and whatever is known in advance. Each of these is a relation that holds only to within a stated tolerance, and the engine weighs them together to find the most probable state of the system, the uncertainty of every quantity in it, and the relative plausibility of competing explanations of the same readings. Each physical domain brings its own laws, units and file formats; the inference and the bookkeeping are shared. Here it carries a chiller: the refrigerant-side heat exchangers, the compressor, the charge inventory and the oil, written as relations among the quantities the sensors report, with one magnitude freed for each candidate fault so that competing explanations of one run can be weighed against one another, and no fault example is ever shown to it. Section 3 states the framework and each of these constructs precisely.

This report is the second of two on that model. The companion report (Marquez 2026) built and fixed everything on 17 normal runs — the laws, the scatter, the priors, an acceptance test of the refrigerant-side model, the identifiability of the faults and a prediction of which fault runs would be detected — and read no fault run. Its acceptance test failed at the strictness fixed in advance, and the protocol had fixed what then follows: the physics-based claims are confined to the two water-flow faults, the only classes whose expression on this machine the acceptance test did not put in doubt, and every comparison between the model and its comparator runs on those runs and the scored normals. This report asks, under that fallback, whether the model names the fault in each of the 31 fault runs without having seen one, how it compares with a rule-based method from the benchmark's own literature, how both degrade as sensors are removed, whether the model's magnitude intervals and stated confidences can be trusted, and whether the detection outcomes agree with what was predicted. It is an inference study: each run's readings, with their stated scatter, and the priors fixed on normal operation are combined into posterior probabilities of the competing hypotheses, from which the answer, the magnitude interval and the stated confidence are all read. Every threshold, population, test and pass rule was fixed in a protocol file committed on 2026-09-22; two dated amendments after the normal-run stage are stated in full in Section 3.11. Section 2 describes the machine, the fault runs and the sensor sets, Section 3 the model, the comparator, the scoring and the protocol, Section 4 the results in the order of the questions, and Sections 5 and 6 what they mean and what they leave open.

2 System and data

The machine is a water-cooled centrifugal chiller of 90 tons nominal capacity, about 316 kW of cooling, with a flooded evaporator, a single-stage centrifugal compressor, a water-cooled condenser with a liquid pool and an integral subcooler, a pilot-operated expansion valve and a heated oil sump. Chilled water carries the cooling load from a test facility into the evaporator and condenser water carries the rejected heat to a cooling tower. The machine, the two loops and every sensor this report uses are shown in Figure 1, which also marks the sensor set at which each reading disappears and where each of the seven faults acts.

The benchmark introduced each fault deliberately at graded severities, each severity as a run over the same grid of loads and water temperatures as the normal runs: condenser tubes plugged by 6 to 45 percent of their number, condenser and evaporator water flows throttled by 10 to 40 percent, refrigerant charge removed and added by 10 to 40 percent of nominal, oil added to the sump by 14 to 50 percent, and nitrogen charged into the condenser at four recorded amounts and one trace amount. Table 1 lists the 31 fault runs this report scores, by family and nominal severity, with the eight normal runs scored for false calls. Each run is a sequence of steady states, 1278 running steady states in all at the laboratory instrumentation after the companion report's point exclusion and duplicate-sensor screen. The truth of a fault run is its family, never its severity: a method is right when it names the family, or the registered class that holds it.

The 17 usable normal runs are the fitting population of every fold and the baseline of every diagnosis. A scored run is paired with the nearest earlier normal run, never across the correction to the condenser-water flow measurement made in September 1999, and the model reads the run's readings as deviations from that paired normal. Eight of the normal runs are scored for false calls as well, each held out with its day block and paired outside it; the other nine are the paired baselines of training fault runs and are reported on but not scored. The hard cases of Table 1 are the mildest quantified severity of four families, which the companion report's identifiability analysis placed inside the scatter of normal operation.

The same measurements are read at five nested sensor sets, from the laboratory instrumentation with 38 channels down to the 5 that a building-automation system without flow meters would hold; Table 2 defines them. They are this report's way of asking which conclusions survive on a real building's instrumentation.

Three facts from the companion report govern what follows and are restated here so that this report stands alone. The acceptance test of the refrigerant-side model failed: its cross-validated error over held-out days was 1.26 and 1.45 times the expected error on the evaporator approach and on the subcooling against a limit of 1.25, with the condenser approach passing at 1.13. On the real normal runs, condenser fouling, non-condensable gas and a refrigerant leak form one identifiability class at their mildest severities ({`cf`, `nc`, `rl`}, {`eo`}, {`ro`}), against the five single families the protocol registered. And the model was predicted to call an untroubled machine faulty in a fraction 0.074 of simulated normal runs with every sensor and about 0.41 without the refrigerant pressures. The model these facts describe is set out as a factor graph in Figure 2: one hypothesis layer above the exchanger state fitted on normal operation, and one reading factor per diagnosis channel and steady state.

Assumptions. The following are held fixed throughout.

- Temperatures are carried in kelvin and powers in watts; severities keep the benchmark's own units, and a flow fault's magnitude is $M = -\ln(1 - x/100)$ for a flow reduced by x percent.

- Every law, prior, noise component and structural range is the one fixed on normal operation; nothing was refitted or retuned after a fault run was read.
- A run's steady states are repeated measurements of one machine state; its answer is one hypothesis for the whole run.
- The registered classes of Table 5 score every method alike; a method is never asked to separate families the instrumentation cannot.
- Normal runs are normal by the benchmark's design, and fault runs carry the fault their name says.

3 Methods

3.1 Overview

The approach diagnoses every scored run once, held out, with a model and a comparator that were both fixed before any fault run was read, and then scores and compares their answers under rules registered in advance. The steps, their inputs and outputs, and the subsection that states each are set out in Figure 3. The normal-run quantities of the companion report and the fault runs of Section 2 are the inputs; each held-out run is diagnosed by the model of Figure 2, stated first as the companion report built it and then as the general probabilistic model it is an instance of, with its rows written out in Section 3.4, and by the comparator; the answers are merged into the registered classes and scored; and the registered comparisons, pass rules, context arms, detection falsifier and sensitivities are computed from the scores. Because the acceptance test failed, every comparison runs on the registered fallback population. The protocol with its amendments and the metrics close the section.

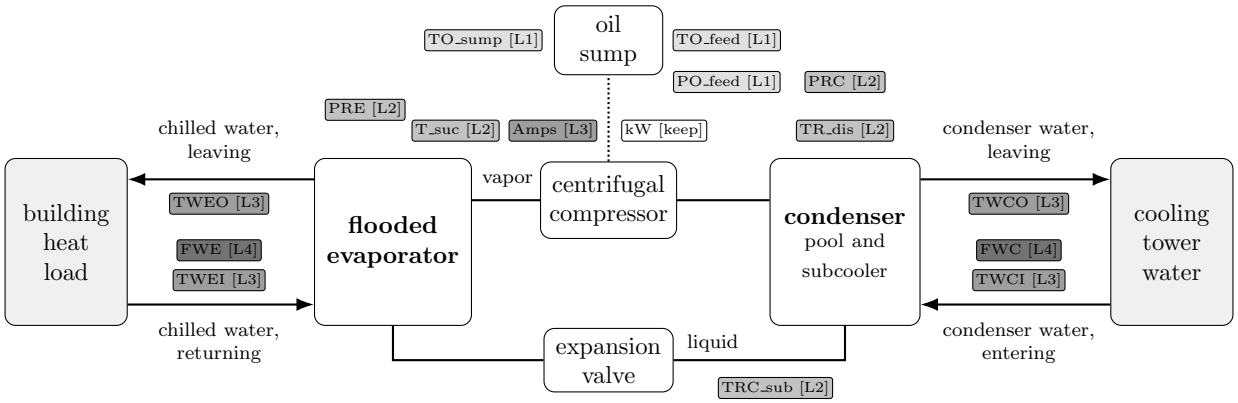
3.2 The diagnosis model

Every scored run is diagnosed held out: the scored runs form 32 groups, each a run with its twins, and each group is held out in turn while the model and the comparator are refitted on the remaining normal runs; the held-out run is then diagnosed at each of the five sensor sets. There are eight hypotheses: normal, and one for each fault family. Each fault frees exactly one magnitude M (dimensionless), constant within a run, carried on a log scale as $u = \ln M$ with a Gaussian prior

$$u \sim \mathcal{N}\big(m_h,\ s_h^2\big) \tag{1}$$

in which $\sim \mathcal{N}(m, s^2)$ reads *is distributed normally with mean m and variance s^2*, and m_h and s_h are the prior mean and standard deviation of the log magnitude under fault hypothesis h. Their values place the severity range the benchmark quantifies inside the prior's bulk, widened by the uncertainty of the map from severity to magnitude. For the two water-flow faults that map is exact; for the other five it depends on constants that normal operation does not identify. Two of them — the share of a nitrogen charge that blankets the tube bundle and the coupling of sump oil into the evaporator — are integrated out of the evidence of the non-condensable and the excess-oil hypotheses rather than fixed; for such a constant ξ,

$$Z_h = \int Z_h(\xi)\, p(\xi)\, d\xi \tag{2}$$



Shading and tag: the sensor set at which a reading disappears. *keep* (white) is present at every set; [L1] the oil instrumentation; [L2] the refrigerant pressures and the refrigerant thermometers; [L3] the four water thermometers on the machine's own nozzles and the compressor current; [L4] the two water-flow meters. Removing a raw sensor also removes every recorded quantity computed from it. The seven faults act where their names say: fouling and non-condensable gas in the condenser, the charge faults in the refrigerant circuit, excess oil in the sump, and the two flow faults in the water loops.

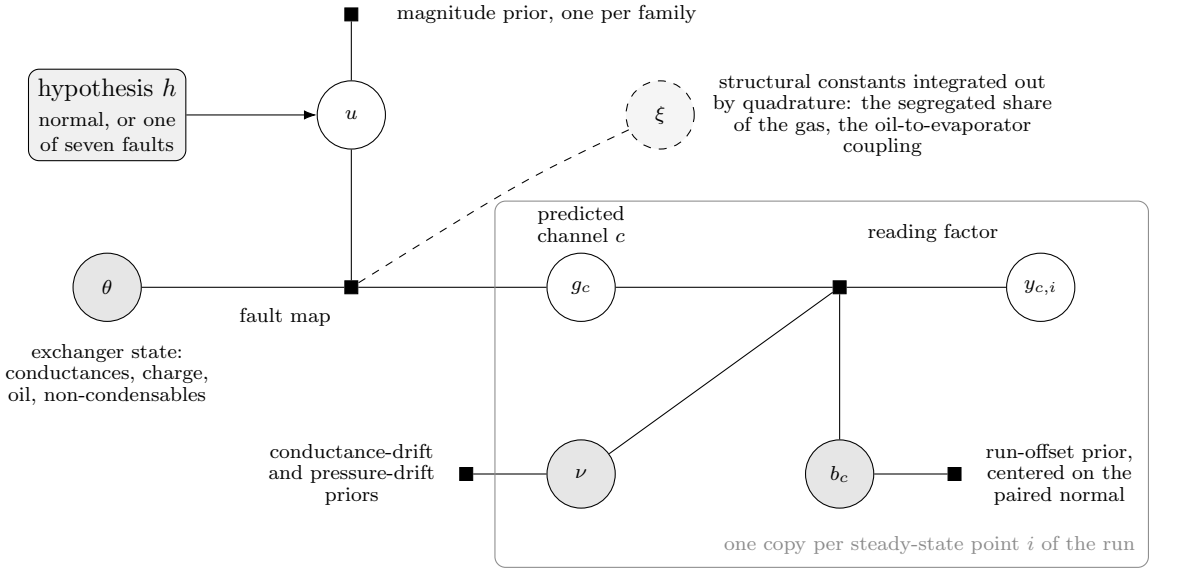
Figure 1: The machine and its two water loops, with every sensor this report uses and the sensor set at which each disappears. Warm return water enters the flooded evaporator, where refrigerant boiling at low pressure cools it; the centrifugal compressor lifts that vapor to the condensing pressure; the condenser rejects the heat to the tower water, and the condensed liquid collects in a pool, is subcooled and expands back through the valve. Lubricating oil circulates from a heated sump through the compressor. The seven deliberately introduced faults are plugged condenser tubes, throttled condenser and evaporator water flows, refrigerant removed and added, oil added to the sump, and nitrogen charged into the condenser.

Table 1: The scored population: 31 fault runs in seven families, each named by its family and its nominal severity, and the eight normal runs scored for false calls (every usable normal run outside the first day block that is not the paired baseline of a training fault run). The four hard cases are the mildest quantified severity of four families; they are scored in the population and also reported on their own. Runs with a letter suffix or a second number are twins of one severity and are held out together.

Family [–]	Fault [–]	Severity in percent of [–]	Runs [–]	Scored runs [–]	Hard cases [–]
`cf`	condenser fouling	tubes plugged	5	`cf6`, `cf12`, `cf20`, `cf30`, `cf45`	`cf6`
`fwc`	reduced condenser water flow	flow reduction	5	`fwc10`, `fwc20 alt`, `fwc20`, `fwc30`, `fwc40`	—
`fwe`	reduced evaporator water flow	flow reduction	4	`fwe10`, `fwe20`, `fwe30`, `fwe40`	`fwe10`
`rl`	refrigerant leak	charge removed	4	`rl10`, `rl20`, `rl30`, `rl40`	`rl10`
`ro`	refrigerant overcharge	charge added	4	`ro10`, `ro20`, `ro30`, `ro40`	—
`eo`	excess oil	excess oil	3	`eo14`, `eo32`, `eo50`	—
`nc`	non-condensable gas	condenser volume	6	`nc trace`, `nc trace2`, `nc1`, `nc2`, `nc3`, `nc5`	`nc trace`
`normal`	normal	—	8	`near normal1`, `near normal2`, `normal r1`, `normal eo`, `normal cf2`, `normal cf3`, `normal cf4`, `normal cf5`	—

Table 2: The five nested sensor sets. Each removes raw sensors from the laboratory instrumentation, and with them every recorded quantity computed from a removed sensor; the channel count is what the methods read at that set.

Set [–]	Instrumentation [–]	Channels [–]	Raw sensors removed, cumulative [–]
L0	the laboratory instrumentation	38	none
L1	no oil instrumentation	33	TO_sump, TO_feed, PO_feed
L2	no refrigerant pressures	22	TO_sump, TO_feed, PO_feed, PRE, PRC, TRC_sub, T_suc, TR_dis
L3	a building-automation point list	13	TO_sump, TO_feed, PO_feed, PRE, PRC, TRC_sub, T_suc, TR_dis, TWEI, TWEO, TWCI, TWCO, Amps
L4	the same without flow meters	5	TO_sump, TO_feed, PO_feed, PRE, PRC, TRC_sub, T_suc, TR_dis, TWEI, TWEO, TWCI, TWCO, Amps, FWE, FWC



The scatter of a reading is split three ways — a day-block component, a run component and a point component — and a discrepancy attached to the operating condition is added to the point component on the corrected channels. **Shaded circles** are shared by every point of a run; **dashed** marks the two structural constants that normal operation does not identify and that are integrated out of the non-condensable and excess-oil hypotheses by quadrature rather than fixed at a midpoint; the condenser water-film share instead widens the condenser-fouling magnitude prior. **The hypotheses are compared by evidence**: for each h the whole graph is fitted and its evidence computed, and the eight evidences are normalized to posterior probabilities. Everything outside the hypothesis layer was fixed on normal operation; this report asks the hypothesis layer its question on the fault runs.

Figure 2: The diagnosis model of one held-out run, as a factor graph. Circles are unknowns and black squares are the relations that tie them. A hypothesis h names one fault family, or normal; it frees a single magnitude, carried on a log scale as u, which a physical fault map turns into a displacement of the exchanger state θ away from the state fitted on normal operation. The displaced state predicts every diagnosis channel of the sensor set; with the laboratory instrumentation these are the evaporator approach, the condenser approach, the subcooling, the oil-sump temperature, the suction superheat, the compressor power, the energy balance, the evaporator water flow and the condenser water flow. Each prediction meets its reading through a factor that also carries a run offset b_c, centered on the run's paired normal, and the run's nuisance vector ν, a conductance drift and a condensing-pressure drift. The structural constants ξ, the segregated share of the gas and the oil-to-evaporator coupling, are integrated out by quadrature; the condenser water-film share enters through the width of the condenser-fouling magnitude prior.

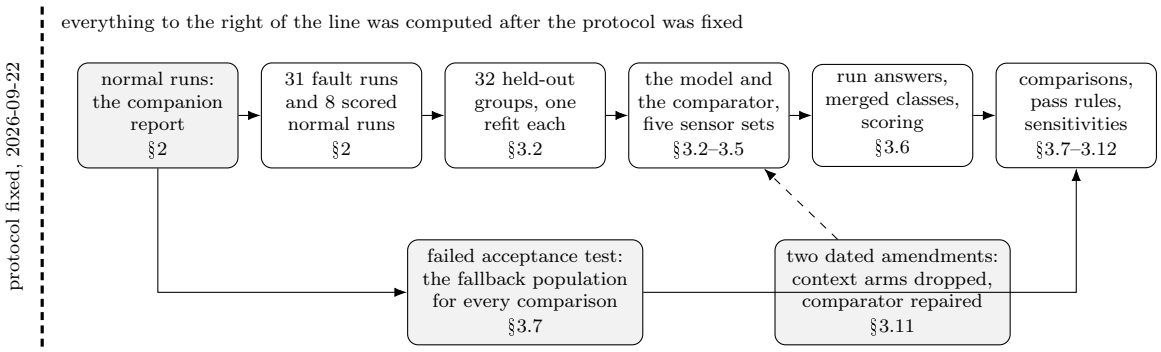


Figure 3: The order of the computation, with the section that states each step marked under it. The companion report's quantities on normal operation — the per-fold models, the verdict of the acceptance test, the identifiability classes and the detection predictions — are read back and never recomputed. Every fault run and scored normal run is held out in turn with its twins, the model and the comparator are refitted on the remaining normals, and each held-out run is diagnosed at five sensor sets. The run answers are merged under each set's registered classes and scored; the comparisons, pass rules and sensitivities follow. Because the acceptance test failed, every comparison with the model runs on the fallback population. Two dated, additive amendments after the normal-run stage dropped the classifier context arms and repaired the comparator; both are described in Section 3.11 and recorded in Table 7.

where $Z_h(\xi)$ is the evidence of hypothesis h with the constant held at the value ξ — the probability of the run's readings under that graph once its free magnitude and nuisances are paid for, computed as in Section 3.3 — and $p(\xi)$ is the constant's prior density over its registered range. The integral is evaluated by quadrature, a weighted sum over a fixed grid of values of ξ, so that those two hypotheses are judged over the whole range of a constant that normal operation leaves open rather than at one guessed value. The other constants widen the magnitude prior instead. The share ψ of the condenser's thermal resistance in the water film sets how much of a plugged-tube area loss the faster flow in the open tubes wins back: with a fraction f_p of the tubes plugged, the magnitude of condenser fouling is

$$M = -\ln \frac{1 - f_p}{\psi\,(1 - f_p)^{0.8} + 1 - \psi} \tag{3}$$

with 0.8 the exponent of the film's conductance in the water velocity, the share held at its central value, and its registered range $[\psi_-, \psi_+] = [0.4, 0.9]$ widening the condenser-fouling prior by the factor its two ends give at the mildest quantified severity f_1, added on the log scale in quadrature to the spread s_{grid} of the severity range,

$$\Gamma_\psi = \sqrt{M(f_1; \psi_-)/M(f_1; \psi_+)} \tag{4}$$

$$s_{\text{cf}}^2 = s_{\text{grid}}^2 + (\ln \Gamma_\psi)^2 \tag{5}$$

and the remaining map constants of the charge, oil and gas families widen their priors in the same way. The fault map displaces the exchanger state fitted on normal operation, and the displaced state predicts every diagnosis channel the sensor set keeps: the evaporator and condenser approaches, the subcooling, the oil-sump temperature and the suction superheat where the refrigerant and oil sensors are present; the compressor power against the lift-coupled power law; the energy balance; and the two water-flow meters where kept. Each reading is the measured value minus the fold's prediction of normal operation at the same operating condition, and a run's offsets b_c are centered on its paired normal's own offset and drawn from the day-block and run components of the scatter

measured on normal operation. Every point of a run enters one joint likelihood with the run's offsets and nuisances — a conductance drift and a condensing-pressure drift — integrated out jointly. The eight evidences of (2) are turned into posterior probabilities of the hypotheses given the run's readings y under equal prior odds, and the run's answer is the most probable one,

$$P(h \mid y) = \frac{Z_h}{\sum_{h'} Z_{h'}} \tag{6}$$

$$\hat{h} = \arg\max_{h} P(h \mid y) \tag{7}$$

where the bar reads *given*, the sum runs over all eight hypotheses h', and $\arg\max_h$ picks the hypothesis with the largest probability. Hypotheses whose probabilities agree to a relative tolerance of 1e-06 are one answer and are ordered by name. Where the sensor set removes the flow meters, the registered design flows replace them in the balance and in the power law and a flow fault scales the balance's duty by e^{-M}.

3.3 The probabilistic model

The layers above give a factor graph, and that graph is solved the same way for every hypothesis. A physical network is a graph, and the engine writes its physics on that graph. The state is a vector z of unknowns: one potential per node, one flow per edge, and whatever else is uncertain, such as demands or parameters. Every relation the state must satisfy is a factor, written as a residual $r_k(z)$ that vanishes when the relation holds and carrying a weight w_k, the inverse variance of how far it may be off. A conservation law at a node, a constitutive law on an edge, a sensor reading and a prior on an approximately known quantity are all factors of this one kind; they differ only in their residual and their weight, a physics relation being trusted to a small tolerance, a reading to its sensor's noise, a prior to its stated spread. Treating each factor as an independent Gaussian belief about its residual gives the posterior density of the state given the readings y,

$$p(z \mid y) \propto \prod_k \exp\!\big(-\tfrac{1}{2}\, w_k\, r_k(z)^2\big). \tag{8}$$

The state that maximizes it, the *maximum-a-posteriori* (MAP) state z^*, minimizes the weighted sum of squared residuals $C(z)$ and is found by Gauss–Newton iteration on the sparse Jacobian J of the residuals, $J_{ki} = \partial r_k / \partial z_i$, with W the diagonal matrix of weights, from a flat start until the gradient norm falls below a tolerance ϵ:

$$C(z) = \tfrac{1}{2} \sum_k w_k\, r_k(z)^2, \tag{9}$$

$$z^* = \arg\min_{z} C(z), \tag{10}$$

$$z \leftarrow z - (J^\top W J)^{-1} J^\top W\, r(z) \quad \text{until } \|J^\top W\, r(z)\| < \epsilon. \tag{11}$$

Near the minimum the cost (9) is approximately quadratic, so the posterior is approximately Gaussian about z^* with information matrix Λ. This is the *Laplace approximation*, and the *posterior standard deviation* σ_i of any unknown is read from the diagonal of Λ^{-1}, from the same sparse factorization the last Gauss–Newton step used:

$$p(z \mid y) \approx \mathcal{N}\big(z^*, \Lambda^{-1}\big), \tag{12}$$

$$\Lambda = J^\top W J, \tag{13}$$

$$\sigma_i = \sqrt{(\Lambda^{-1})_{ii}}. \tag{14}$$

When several factor graphs compete to explain the same readings — a leak here, a leak there, no leak — each is compared by its *evidence* $Z = \int \prod_k \exp(-\frac{1}{2} w_k r_k(z)^2)\, dz$, the probability of the readings under the model once the extra freedom of its unknowns has been paid for. In the Laplace approximation the evidence has a closed form, and normalizing every factor by its own weight lets graphs with different numbers of unknowns d and factors compare as densities,

$$\log Z = -C(z^*) + \tfrac{d}{2} \log 2\pi - \tfrac{1}{2} \log\det \Lambda + \sum_k \tfrac{1}{2} \log \frac{w_k}{2\pi}. \tag{15}$$

Differences of $\log Z$ between hypotheses, from (15), are in *nats*, natural-log units of odds; five nats is odds of about 150 to 1. The posterior probability of a hypothesis is its evidence normalized over all hypotheses under equal prior odds. Everything that follows is a choice of variables and factors on this graph, followed by these three operations.

The model, the fit and its uncertainty. The quantities to be determined are the fault magnitude of the hypothesis under test, the run's offsets on every diagnosis channel, and its drift nuisances. Physics ties them together and the sensors constrain them. At every node a conserved quantity balances: at every steady state the heat the condenser rejects equals the evaporator duty plus the compressor power, to within the balance's scatter. The constitutive laws — each heat exchanger's conductance, which ties its approach temperature to its duty and its water flow, and the compressor's efficiency law, which ties its power to the refrigerant lift and the load — are not rows of the graph: they are evaluated in closed form inside each channel's predicted shift (Section 3.4), and the only closure row is the flow term of a water-flow fault whose meter the sensor set removes. Every sensor adds one relation between its reading and the quantity it observes (each diagnosis channel's deviation from the fold's prediction of normal operation equals the hypothesis's predicted shift plus the run offset, centered on the paired normal, to within the point scatter), and every quantity known approximately in advance adds a prior, a relation between the unknown and its expected value (the magnitude prior of (1), the offset prior centered on the paired normal, and the nuisance priors). Each relation is written as a residual that vanishes when it holds, and together the residuals form one sparse system; this report calls them rows. No row is enforced exactly. Each carries a weight, the inverse variance of how far it may be off: a physics row is trusted to a small tolerance, a reading to its sensor's noise, a prior to its stated spread. The *maximum-a-posteriori* (MAP) state is the set of unknowns that minimizes the weighted sum of squared row misfits, $C(z) = \frac{1}{2} \sum_k w_k\, r_k(z)^2$; with only physics and readings it is the weighted least-squares fit, and it is found by Gauss–Newton iteration. The *posterior standard deviation* of an unknown is its uncertainty bar: near the minimum the misfit is approximately quadratic, so the state is approximately Gaussian with information matrix $\Lambda = J^\top W J$ (the weighted Jacobian of the rows), and the standard deviation of an unknown is the square root of the corresponding diagonal entry of Λ^{-1}. Treating the state as Gaussian around the MAP is the *Laplace approximation*. When several models compete — a leak here, a leak there, no leak — each is compared by its *evidence*, the probability of the readings under the model once its extra freedom is paid for: in the Laplace approximation

$$\log Z = -C(z^*) + \tfrac{d}{2} \log 2\pi - \tfrac{1}{2} \log\det \Lambda + \sum_k \tfrac{1}{2} \log \frac{w_k}{2\pi},$$

with d the number of unknowns and the last sum over every row so that models of different size compare as densities. Differences of $\log Z$ are in *nats* (natural-log units of odds; 5 nats is odds of

about 150 to 1), and the models' posterior probabilities are their evidences normalized to sum to one under equal prior odds.

3.4 The residual rows of one hypothesis

The graph of Section 3.3 is assembled once for each run, sensor set and hypothesis, and its rows are written out here in full; they are those of the companion report, assembled on the fault runs. A run has I steady-state points, indexed by i. Each reading enters as its deviation from the fold's prediction of normal operation at that point's own operating condition, and the energy balance enters as the relative imbalance of the chiller's heat flows,

$$\tilde{y}_{c,i} = y_{c,i} - \hat{y}_{c,i} - \delta_{c,t_i}, \tag{16}$$

$$\tilde{y}_{E,i} = \frac{Q_{C,i} - P_i}{D_i} - 1, \tag{17}$$

where $y_{c,i}$ is the reading of channel c at point i, $\hat{y}_{c,i}$ the fold's fitted laws evaluated at that point's water temperatures, flows and load, and δ_{c,t_i} the per-condition discrepancy at the point's operating condition t_i, the training normals' mean residual there, which is zero on the two flow meters; D_i is the evaporator duty, $Q_{C,i}$ the heat the condenser rejects and P_i the compressor power, all in kilowatts. A flow meter's deviation is its reading less the fold's nominal flow, and the suction superheat is read as its excess over the dry zone's normal share of the measured evaporator approach. The run's offsets b_c, one per channel with b_E the balance's, its nuisance components ν_k and, under a fault, the log magnitude u are the unknowns.

The rows follow the physics. The one conservation row per point is the chiller's energy balance per unit evaporator duty, the condenser heat less the power less the duty, written against the run's offset b_E on the imbalance (17),

$$r_{E,i} = b_E - \tilde{y}_{E,i} + \mu_i, \tag{18}$$

where μ_i is present only when the hypothesis is a water-flow fault on a loop whose meter the sensor set removes. The balance then runs on the registered design flow, and the flow term carries the reduction through one closure row per point,

$$r_{F,i} = \mu_i - (1 + b_E)\big(s_E(u) - 1\big) + q_i\big(s_C(u) - 1\big), \tag{19}$$

with $q_i = Q_{C,i}/D_i$, $s = e^{-M}$ on the faulted loop and one on the other, and a spread 0.0001 times that of the balance, so that (19) defines μ_i without stiffening any other unknown; $\mathbf{1}_F$ is one in such a graph and zero otherwise. Each reading channel c of the set, the balance excluded, adds one measurement row per point,

$$r_{c,i} = \tilde{y}_{c,i} - b_c - \sum_k G_{c,i,k}\,\nu_k - g_{c,i}(u), \tag{20}$$

where $g_{c,i}(u)$ is the hypothesis's predicted shift of the channel at the point, zero under normal, and $G_{c,i,k}$ is the response of the channel at the point to one unit of nuisance component k: the condensing-pressure ramp, a centered linear ramp over the run's points on the condenser approach and the subcooling, and the components of the conductance drift, the model's own response to a shift of each loop's $\ln UA$. The row's spread is the channel's point standard deviation s_c on the fold model's residuals, widened on the channels the discrepancy corrects by the error of the condition means,

$$\tilde{s}_c^2 = s_c^2 + \frac{1}{I}\sum_i \frac{s_c^2}{n_{t_i}}, \tag{21}$$

with n_t the number of training normals at condition t; at a condition none of them visited, the variance of a single training point there replaces s_c^2/n_t. The balance row's spread is the balance's own point standard deviation s_E. The prior rows close the graph, one per unknown other than the flow terms,

$$r_{b,c} = b_c - m_c, \tag{22}$$
$$r_{\zeta,j} = \zeta_j, \tag{23}$$
$$r_{\nu,k} = \nu_k, \tag{24}$$
$$r_u = u - m_h, \tag{25}$$

where the first is the offset prior of a channel outside the joint block — the balance, the meters, and the power where no exchanger channel is kept — and the second is the whitened form of the joint offsets of the exchanger channels and the power, whose prior covariance $V_b = LL^\top$ correlates them,

$$b_c = m_c + \sum_j L_{cj}\,\zeta_j. \tag{26}$$

A nuisance component is carried at unit spread, its scale being in $G_{c,i,k}$, and the magnitude row is the prior of (1). The offset prior is centered on the paired normal's own offset m_c, and for a fold with n_B training day blocks its variance is

$$\sigma_{b,c}^2 = \frac{n_B - 1}{n_B - 3}\Big(2\big(s_{B,c}^2 + s_{R,c}^2\big) + \frac{s_c^2}{I_0}\Big) - \omega_c, \tag{27}$$

with $s_{B,c}$ and $s_{R,c}$ the day-block and run standard deviations on the fold model's residuals, I_0 the paired normal's number of points, and ω_c the between-block variance the conductance drift carries in its place; the joint block carries the matrix form of the same expression, with the variance of the oil-service shift added on the sump, and every offset's spread is floored at a small fraction of s_c. Both depend on the fold and the run, and are not tabulated.

Table 3 lists every row family with its residual, unit, spread, weight and multiplicity, and Table 4 gives the multiplicities at each sensor set with the point counts of the diagnosed runs. For hypothesis h at set s the residual vector r, the Jacobian J and the weight matrix W of (11) have the dimensions

$$\dim r = I\,\big(1 + |\mathcal{C}_s| + \mathbf{1}_F\big) + n_p, \tag{28}$$
$$n_p = n_b + |\mathcal{J}_s| + n_\nu + \mathbf{1}[h \neq \text{normal}], \tag{29}$$
$$\dim z = n_p + I\,\mathbf{1}_F, \tag{30}$$

so that J is $\dim r$ by $\dim z$ and W is the diagonal $\dim r$ by $\dim r$ matrix of the weights; n_p counts the prior rows, $\mathbf{1}[\cdot]$ is the indicator, and the $\dim z$ unknowns are one per prior row and one flow term per point of a flow-term graph. Across the sensor sets the row families stay the same and their multiplicities change: a removed sensor takes with it the reading rows of every channel it feeds, the joint block goes with the exchanger channels and the ramp with the refrigerant pressures, and at the set without flow meters the water-flow faults gain the closure rows of (19). Without the refrigerant pressures the power row's prediction comes from the lift rebuilt from design approaches, and its spread widens accordingly.

Some families a reader of Figure 2 might expect are absent. The graph contributes no closure row for the exchanger or the compressor: the conductance laws, the charge inventory, the oil law

and the power law are evaluated in closed form inside $\hat{y}_{c,i}$ and $g_{c,i}(u)$, and the derivative of g in u is a central difference with step 10^{-5}. The energy balance is the only conservation row; the refrigerant side carries no balance of its own. There is no boundary or control row, because the water temperatures, flows and load of every point enter as known inputs, and there is no discrepancy, regularizing or ridge row: the per-condition discrepancy is a fixed correction δ and a variance term in (21), not an unknown, and the fit runs without a ridge. For excess oil and for non-condensable gas the graph is assembled once per quadrature node of its structural constant, five nodes each, with the same rows and only $g_{c,i}(u)$ changing; the other families carry their structural spread in s_h. A registered sensitivity that sets the ramp's span to zero assembles the same graph without the ramp row.

Several calculations of this report lie outside the graph and contribute no row to it: the per-fold fits of the laws, each by unweighted least squares on the training normals; the scatter components and the condition means δ, estimated by moments on the same normals; the comparator of Section 3.5, its regressions and thresholds; the merging into classes, the scores and the paired comparisons of Section 3.6 and Section 3.7; and the context arms and the detection falsifier. The governing evidence of each hypothesis is computed on these rows by integrating the offsets and nuisances exactly, since every row is linear in them, and u by quadrature on a grid of 200 points. Beside it, the maximum-a-posteriori solve of (11) and the Laplace evidence of (15) are computed on the same rows for every run, set and hypothesis, one solve per quadrature node where a constant is integrated out; across the 240 run diagnoses at the scored centering the most probable hypothesis they give differs from the governing one on none of them.

3.5 The comparator

The comparator is the five-feature regression method that Reddy (2007) ranked best of four on this benchmark, reproduced from its description and refitted here on each fold's training normals only, so that none of its author's in-sample evaluation carries over. Table 6 states its configuration. Five characteristic features are computed from the workbook channels in workbook units — the evaporator water temperature difference, the condenser water temperature difference, the subcooling, the condenser approach, and the condenser conductance

$$\mathrm{CP1} = c_p\, \dot{m}_{cd}\, \ln \frac{T_{rc} - T_{ci}}{T_{rc} - T_{co}} \tag{31}$$

with T_{rc} the condensing temperature, T_{ci} and T_{co} the entering and leaving condenser water temperatures, c_p the specific heat of water and $\dot{m}_{cd}$ the condenser water mass flow; it is the log-mean-temperature-difference form of an exchanger's conductance, the heat rejected divided by the mean temperature difference between refrigerant and water, and it assumes a condensing temperature uniform along the tubes — and each feature is given a least-squares reference model in the chilled-water leaving temperature, the condenser-water entering temperature and the load, with a fixed second-order term list. The comparator flags a point by how far each feature departs from its reference model, in units of that model's scatter: the standardized residual on feature k is

$$t_k = \frac{y_k - \hat{y}_k}{\mathrm{RMSE}_k} \tag{32}$$

where y_k is the feature at the point, $\hat{y}_k$ the reference model's prediction at the same operating condition and RMSE_k that model's root-mean-square error on the training normals, so t_k is dimensionless. The point is flagged when $\max_k |t_k|$ exceeds one threshold c, set by leave-one-day-block-out cross-fitting on the training normals as the smallest order statistic that leaves at most the registered

Table 3: The complete residual ledger of one hypothesis h on one run of I points at sensor set s: every row the graph assembles, grouped conservation, closure, measurement, prior; the graph has no boundary, control, discrepancy or regularizing row. *Residual* is the row's r_k with its sign, from (18), (19), (20) and (22). σ_k is the spread of the row, in the row's unit, and $w_k = \sigma_k^{-2}$ its weight; the values are those of the laboratory instrumentation (L0) in the fold trained on every usable normal, and each held-out fold re-estimates them on its own training normals. A reading row's spread is $\tilde{s}_c$ of (21); the column gives its point part s_c and s_c^{-2}. *Rows* is the multiplicity: $\mathcal{C}_s$ is the set's reading channels, $\mathcal{J}_s$ its joint-offset channels, n_b its other offsets and $\mathbf{1}_F$ one in a flow-term graph, all counted in Table 4; a nuisance has as many rows as its prior covariance has nonzero directions, at most the number shown, and the magnitude row is present under a fault only. In *Source or role*, *ramp* and *drift* name the nuisances that reach a reading channel, the condensing-pressure ramp and the conductance drift, and the closure row is assembled for the water-flow faults at the sensor set named.

Row [–]	Residual r_k [unit]	Unit [–]	σ_k, L0 [unit]	w_k, L0 [unit^{-2}]	Rows [count]	Source or role [–]
Conservation						
E_i	$b_E - \tilde{y}_{E,i} + \mu_i$	–	0.0320569	973.099	I	energy balance per unit duty
Closure						
F_i	$\mu_i - (1 + b_E)(s_E - 1) + q_i(s_C - 1)$	–	3.20569×10^{-6}	9.73099×10^{10}	$I\,\mathbf{1}_F$	defines μ_i; {`fwc`, `fwe`} at L4
Measurement						
$Y_{c,i}$	$\tilde{y}_{c,i} - b_c - \sum_k G_{c,i,k}\nu_k - g_{c,i}(u)$				$I\,\lvert\mathcal{C}_s\rvert$	per point and channel
TEA	as $Y_{c,i}$	K	0.158385	39.8632	I	evaporator approach; drift
TCA	as $Y_{c,i}$	K	0.193977	26.5766	I	condenser approach; ramp, drift
TRC_sub	as $Y_{c,i}$	K	0.195304	26.2167	I	subcooling; ramp, drift
TO_sump	as $Y_{c,i}$	K	0.282864	12.4981	I	oil sump; drift through the oil law
Tsh_suc	as $Y_{c,i}$	K	0.0926041	116.611	I	superheat over the dry zone's normal share
power	as $Y_{c,i}$	kW	0.567034	3.11015	I	power against the lift law; drift
FWE	as $Y_{c,i}$	L/s	0.0313649	1016.51	I	meter less nominal flow; `fwe` only
FWC	as $Y_{c,i}$	L/s	0.109755	83.014	I	meter less nominal flow; `fwc` only
Prior						
B_c	$b_c - m_c$	of c	$\sigma_{b,c}$	$\sigma_{b,c}^{-2}$	n_b	other offsets, (27)
H_j	ζ_j	–	1	1	$\lvert\mathcal{J}_s\rvert$	joint offsets, (26)
N_{p}	ν_{p}	–	1	1	≤ 1	ramp, span sd 0.1 K
$N_{\mathrm{d},k}$	$\nu_{\mathrm{d},k}$	–	1	1	≤ 2	drift, sd floor 0.02 in $\ln UA$
U	$u - m_h$	–	s_h	s_h^{-2}	1, fault	magnitude prior (1)

Table 4: The multiplicities of Table 3 at each sensor set, read from the fold trained on every usable normal, and the diagnosed runs. $|\mathcal{C}_s|$ counts the reading channels, $|\mathcal{J}_s|$ the channels whose offsets share the joint prior of (26), n_b the offsets outside it, the balance's among them, and n_ν the nuisance rows at most. *Rows per point* is $1 + |\mathcal{C}_s|$, the balance row and the reading rows, and a flow-term graph adds one more. I is the range of the diagnosed runs' point counts and *Runs* their number, the fault runs and every normal run; (28) gives the rows of one graph. *Offset priors* is $n_b + |\mathcal{J}_s|$; (29) adds the nuisance rows and, for a fault, the magnitude row. The last two columns are the point spreads of the balance row and of the power reading row.

Set [–]	$\lvert\mathcal{C}_s\rvert$ [count]	$\lvert\mathcal{J}_s\rvert$ [count]	n_b [count]	n_ν [count]	Rows per point [count]	Flow-term faults [–]	I [count]	Runs [count]	Offset priors [count]	s_E [–]	s_c, power [kW]
L0	8	6	3	≤ 1	9	—	19–27	48	9	0.0320569	0.567034
L1	7	5	3	≤ 3	8	—	19–27	48	8	0.0320569	0.567034
L2	3	0	4	≤ 2	4	—	19–27	48	4	0.0320569	1.07457
L3	3	0	4	≤ 2	4	—	26–27	48	4	0.0321964	1.07273
L4	1	0	2	≤ 2	2	{`fwe`, `fwc`}	26–27	48	2	0.0332907	1.0834

fraction 0.05 of held-out normal points above it. A flagged point is isolated by the author's table of expected symptom signs: a rule fires when every cell it constrains matches the sign of (32), and the point's answer is the one fault whose rule fired, or ambiguous when several or none fired. The run's answer is the plurality of its points' answers, normal winning a tie and ambiguous losing every tie. Three cells of the published table are not fixed by the accessible text and are resolved by the reading most consistent with its logic; the alternatives are registered sensitivities, as are two further false-alarm rates and a relaxed match in which two of three symptoms suffice. The method has no rule for excess oil; the oil-sump rule of Zhou et al. (2009) is run as a separately labeled extension and never stands in for it.

Two facts about the comparator are disclosed because they bear on the comparison. Its author derived the sign table from the benchmark's most severe fault runs, so the comparator carries design-time label information from the very runs it is scored on, a leak that favors it; the model reads no fault run at any stage. And the conductance of (31) has no value at a steady state whose condensing temperature is not above both condenser water temperatures, which the refrigerant cannot do but a sensor can report; the published data holds such rows, the comparator as first built stopped with an error on them, and the repair of Section 3.11 makes it set such rows aside and count them. Where a sensor set removes a feature's channel that feature leaves detection and every rule; where it removes a regression input the method cannot run and every answer is not applicable, scored wrong.

3.6 Scoring and the merged classes

Every method is scored on the same 39 runs at every sensor set: the 31 fault runs, whose truth is their family, and the eight scored normals, whose truth is normal. The unit is the run, and run-level accuracy carries an exact binomial 95 percent interval (Clopper and Pearson 1934). Two scorings are published side by side. The unmerged scoring treats each family as its own class. The merged scoring uses the classes of Table 5, registered as values before any fault run was read from the model's own fault signatures in the channels each set keeps: at the laboratory instrumentation every family is its own class, so the two scorings coincide; where the oil sensor is removed, fouling, the leak and excess oil form one class; and where the refrigerant pressures are removed the overcharge joins them, since at those sets every refrigerant-side fault reaches the diagnosis through the compressor

lift alone. The same classes score the comparator: its ambiguous answer is credited only when every rule it fired lies in one class that holds the truth. Merged scoring removes the model's own structural confusions by construction, which is why the unmerged count is printed beside every merged one. An analysis that fails is the answer *failed*, scored wrong and counted; none failed.

Table 5: The classes a method's answer is scored against at each sensor set. The registered classes were fixed as values before any fault run was read, from the model's own fault signatures in the channels each set keeps; the two water-flow families and normal are always classes of their own. The last column lists any further families the companion report's visibility rule tied at that set, which the scoring also merges; none were tied.

Set [–]	Registered classes among the five refrigerant-side families [–]	Families tied by the model [–]
L0	{`cf`}, {`eo`}, {`nc`}, {`rl`}, {`ro`}	none
L1	{`cf`, `eo`, `rl`}, {`nc`}, {`ro`}	none
L2	{`cf`, `eo`, `rl`, `ro`}, {`nc`}	none
L3	{`cf`, `eo`, `rl`, `ro`}, {`nc`}	none
L4	{`cf`, `eo`, `rl`, `ro`}, {`nc`}	none

Table 6: The configuration of the model-based comparator, fixed in the protocol. The faithful method is the one in the confirmatory family; every sensitivity and the excess-oil extension are reported apart and never stand in for it.

Setting [–]	As registered [–]	Registered sensitivities [–]
False-alarm rate of the threshold	0.05	0.025, 0.1
Reading of the ambiguous table cells	primary	fwe on condenser dt, cf without approach, cf without ua
Symptom match	every constrained cell	two of three
Excess-oil rule	not Reddy's; run apart	—
Calibration groups of the threshold	normal day block	—
Duplicate-sensor screen	applied	—

3.7 The comparisons and the pass rules

One family of comparisons is confirmatory. The model's merged run-level answer is set against the comparator's at four sensor sets — the comparator cannot run at the fifth — with the exact conditional McNemar test (McNemar 1947) on the discordant runs, those exactly one method answers correctly:

$$p = 2\,\min\Big(\sum_{j\le n_{01}} \binom{n_{10}+n_{01}}{j} 2^{-(n_{10}+n_{01})},\ \sum_{j\ge n_{01}} \binom{n_{10}+n_{01}}{j} 2^{-(n_{10}+n_{01})}\Big) \tag{33}$$

with n_{10} the runs only the model answers right and n_{01} the runs only the comparator does, and $p = 1$ when there is no discordance. The difference in merged accuracy Δ, model minus comparator, carries a percentile bootstrap interval (Efron 1979) over the 32 held-out groups, twins together, with 10,000 draws from one shared draw matrix for the whole family. The four p-values are corrected by Holm's step-down procedure (Holm 1979) at a family-wise level of 0.05, and the model is stated to outperform the comparator at a set only if Holm rejects, the interval lies on the model's side of zero and the discordant runs favor the model; otherwise the verdict is *not distinguishable at this sample size*. Because the acceptance test failed, every comparison with a model arm runs on the

fallback population of 17 runs — the nine water-flow fault runs and the eight scored normals — and Table 14 restates what that population can resolve, as registered before any verdict was read.

Three further quantities carry registered pass rules or are reported without a test. Detection at the registered operating points compares the model's *answer is not normal* with the comparator's flagging, by family and severity, with each method's false calls on the scored normals; it is descriptive. Magnitude coverage asks whether the model's 95 percent magnitude interval under the true hypothesis contains the nominal magnitude; on the two water-flow families, whose map is exact, the rule passes at a set when the exact binomial interval of the coverage includes 0.95, and the claim holds when it passes at L0 and L1. Calibration asks whether the mean stated confidence of the most probable merged class lies inside the exact binomial interval of that class's accuracy over the scored runs; the claim holds when it passes at L0 and L1, and the Brier score, the log score and a reliability diagram are reported at every set. The degradation curve plots each method's merged accuracy against the sensor set and carries no test. Every other number in this report — the sensitivities, the context arms, the comparator's variants, the hard cases — is descriptive and is never worded as a finding.

3.8 Context arms without labels

Three label-free arms are run on the model's own residual rows, so that the model's answers can be read against simpler uses of the same residuals. Two are detectors at a five percent false-alarm rate: Hotelling's T^2 bound on the Gaussian fitted to the training normals' residual rows (Hotelling 1931), and a one-class support-vector machine with a radial kernel (Schölkopf et al. 2001); a run is marked when its fraction of marked rows exceeds the 95th percentile of the training normals', cross-fitted. The third names a family: the directional arm compares the run's mean residual with each family's expected signature at its lowest severity, read from the companion report, by the cosine of the angle between them, and a sign rule table on the same residuals with a deadband of three training-normal standard deviations is reported beside it. The protocol also registered sixteen label-trained classifier arms with their leakage controls as context; the amendment of Section 3.11 dropped them, so the question of a label-trained comparator is answered by the cited literature only.

3.9 The detection falsifier

The companion report predicted, for each fault run and sensor set, the probability that the model would answer *not normal*, from 60 simulated draws at the run's recorded severity, and marked each family's lowest and highest quantified severity as expected detectable or not. Outcomes that follow from the instrumentation alone — a flow fault read directly by its own flow meter, the deepest excess-oil severity read directly by the oil sump — were set aside by name in advance and never count as confirmations or contradictions. A run contradicts its prediction when its outcome had probability below 0.05 under the prediction; and the number of runs detected is compared with the sum of the predicted probabilities by the Poisson-binomial two-sided test.

3.10 Sensitivities

Every structural input and prior the companion report registered with a range is moved to each end of that range in turn — the charge inventory's overfeed share and superheat authority, the condensing-pressure drift's span, the conductance drift at the covariance the fold supports, the non-condensable gas moving the conductance rather than the pressure, every structural prior width at half and at twice its value, and at the lower sets the design approaches and design flows — and the

number of the 48 diagnosed runs whose most probable hypothesis changes is counted. The model is also run with the offset centered on the fold's mean normal rather than on the paired normal, and the scored normals are re-paired with a partner from their own day block. A pooled score with fouling and non-condensable gas as one class is reported for the laboratory instrumentation, where the companion report found the two inseparable at their mildest severities. None of these governs any claim.

3.11 Protocol, registration and amendments

Every number above was fixed in a file committed on 2026-09-22, before any component of the method had read a fault run (Nosek et al. 2018), and the computation ran in two stages: the normal-run stage of the companion report, committed alone, and the fault-run stage of this report, which reads the normal-run quantities back and applies the acceptance verdict and the identifiability classes mechanically. A chain of checks, run offline on the repository's history, holds that the registration precedes the normal-run results, that these precede the first fault-run results, that no commit between the software each stage names and its results touches the model, the comparator or the stages, and that the values of the earlier files never change afterwards. Table 7 carries the commits.

Three dated, purely additive amendments were made to the protocol file, each naming the software it was written against. The first, on 2026-09-22, preceded the normal-run stage and changed how undefined quantities are written out; the companion report describes it. The two that follow were made after the normal-run stage and before any fault-run value existed. On 2026-11-08, the nine registered blocks that train a classifier on fault labels were dropped from the fault-run stage: timed at one arm each, they were nearly the whole of the stage's compute — more than a day of wall time against minutes per sensor set for the model — and they had already been demoted to context when the confirmatory family was fixed. They were dropped rather than resized so that nothing kept would change; Table 8 lists them, and the results file carries a marker at each. The consequences for this report are that no classifier number appears, that the registered leakage control on the classifier literature is not reported, and that the question of a label-trained comparator is answered by citation only. On 2026-11-08, a first fault-run computation under that deviation finished with every configuration of the comparator failing at the two sensor sets that keep the refrigerant channels: one steady state of one training normal has a condensing temperature below the leaving condenser water, where (31) has no value, and the comparator stopped with an error on it in every fold whose training set held that run. The comparator was repaired to set such rows aside and count them, the amendment records exactly what had been seen of the failed computation — its failure list, and a reproduction of one fold's fit that printed the error, the row counts and the calibration thresholds — and the stage was run again. No answer, score, p-value or verdict of the failed computation was read. The screen set aside one training steady state (normal cf, test 23) and one point each of `fwe30` and `normal cf` at L0 and L1, and no row at the three lower sets.

3.12 Metrics

The measures are grouped by the question of Section 1 each answers, and each is read against a stated reference. All are dimensionless, the log score being stated in nats.

Naming the fault. Over the N scored runs at a sensor set, let τ_r be the true family of run r (normal for a normal run) and $\hat{H}_r$ its answer as a registered class of Table 5, a set of one or more families; for the model it is the class holding the hypothesis of (7). The merged run-level accuracy

is the fraction of runs whose answer holds the truth, and the comparison between two methods is the difference of their accuracies,

$$A = \frac{1}{N} \sum_{r=1}^{N} \mathbf{1}\big[\tau_r \in \hat{H}_r\big] \tag{34}$$

$$\Delta = A_{\text{model}} - A_{\text{comparator}} \tag{35}$$

where the indicator $\mathbf{1}[\cdot]$ is one when the statement in brackets holds and zero otherwise, and $\in$ reads *is a member of*. With every family its own class, (34) is the unmerged accuracy. A is quoted with the exact binomial 95 percent interval of Clopper and Pearson (1934), the range of success probabilities under which the observed count lies in neither 2.5 percent tail, and is read against the comparator's at the same set and against the unmerged count beside it; Δ is read against zero, with its bootstrap interval and the p-value of (33) read against Holm's threshold.

Detection and false calls. With n_F fault runs and n_0 normal runs counted, and $\hat{h}_r$ a method's answer on run r, the detection rate and the false-call rate are the fractions of fault runs and of normal runs answered as anything but normal,

$$d = \frac{1}{n_F} \sum_{r \in \text{fault}} \mathbf{1}\big[\hat{h}_r \neq \text{normal}\big] \tag{36}$$

$$f = \frac{1}{n_0} \sum_{r \in \text{normal}} \mathbf{1}\big[\hat{h}_r \neq \text{normal}\big] \tag{37}$$

where each sum runs over the runs of the kind named under it. A perfect method has $d = 1$ and $f = 0$; f is read against the false-call rate the companion report predicted for each set from simulated normal operation, and d is descriptive.

Magnitude. For the n water-flow fault runs, let $[l_r, u_r]$ be the model's 95 percent posterior interval of the magnitude under the run's true hypothesis and $M_r = -\ln(1 - x_r/100)$ the nominal magnitude of a flow reduced by x_r percent. The coverage is the fraction of intervals that contain the nominal,

$$\kappa = \frac{1}{n} \sum_{r=1}^{n} \mathbf{1}\big[l_r \leq M_r \leq u_r\big] \tag{38}$$

which a calibrated 95 percent interval gives at about 0.95; the rule passes at a set when the exact binomial interval of κ includes 0.95.

Confidence. Let $\tilde{H}_r$ be the merged class the model makes most probable for run r and π_r the probability it gives that class. The mean stated confidence and the accuracy of the top class are

$$\bar{\pi} = \frac{1}{N} \sum_{r=1}^{N} \pi_r \tag{39}$$

$$a_{\text{top}} = \frac{1}{N} \sum_{r=1}^{N} \mathbf{1}\big[\tau_r \in \tilde{H}_r\big] \tag{40}$$

and a calibrated model, one whose stated probabilities are borne out as frequencies, has $\bar{\pi}$ close to a_{top}; the rule passes at a set when $\bar{\pi}$ lies inside the exact binomial interval of a_{top}. Two scores weigh the whole posterior rather than its top class. With $P_r(K)$ the model's probability of merged

class K for run r, the Brier score is the mean squared distance between the posterior and the truth, summed over the classes K of the set,

$$B = \frac{1}{N} \sum_{r=1}^{N} \sum_{K} \left(P_r(K) - \mathbf{1}[\tau_r \in K] \right)^2 \tag{41}$$

which is 0 for a method certain and right on every run and 2 for one certain and wrong on every run; and the log score is the mean of the negative natural logarithm of the probability given to the class holding the truth, K_r,

$$S = -\frac{1}{N} \sum_{r=1}^{N} \ln P_r(K_r) \tag{42}$$

in nats, 0 when the truth is always given probability one and without bound when it is given probability zero. Both are read across the sensor sets, lower being better.

Detection predictions. With p_r the probability of detection the companion report simulated for run r, the expected number of detected runs among the n predictable runs is

$$E = \sum_{r=1}^{n} p_r \tag{43}$$

and the observed count is read against it by the Poisson-binomial test: the exact distribution of a count of independent yes-or-no outcomes with unequal probabilities, from which a two-sided p-value follows.

Table 7: The record that fixes what was decided before each computation ran. The protocol file was committed on the date shown; the software it names was committed before it; each amendment names the software it was written against (its own commit's parent); and each computation ran at the commit shown with no change to the model, the protocol or the measurements in between. The two amendments after the normal-run computation are the deviation and the repair of Section 3.11.

What [–]	Date [–]	Commit [–]
Protocol registered	2026-09-22	`b3e770b127659e3dbb9875f13da4403b1a3a4357`
Software the protocol named	2026-09-22	`6268dd93cea70a91559e0e9857d11917d573d922`
Software correction, before normal operation was computed	2026-09-22	`0aeb14c4e7571737a37aa0a29fff8c0f0feec060`
Computation on normal operation	—	`01faf6d5d538e172e7eb850c0ff5070af426250d`
Protocol deviation, after normal operation was computed	2026-11-08	`090eb3d582f62d574dd12820cfe6a3bdc82b352b`
Comparator repair, after a failed first computation	2026-11-08	`fe5f7b349c31e10a60ace15d81f6e3f18ed99475`
Computation on the fault runs	—	`9fdfd4a61c5e6d53e416079c8045305bdef4e0d4`

4 Results

4.1 Diagnosis accuracy across the sensor sets

Every method's run-level accuracy of (34) at every sensor set, merged and unmerged, is set out in Table 9, one row per method and set with the count, the rate and its exact interval; the two confirmatory methods' merged accuracies are plotted against the set in Figure 4, one mark per method and set with its interval as a vertical bar. With the laboratory instrumentation the model

Table 8: The nine blocks the deviation amendment of 2026-11-08 dropped from the fault-run computation. Every one trains a classifier on fault labels; none feeds a number of the model, of the comparator or of any comparison between them. The results file carries a marker at each path naming the amendment.

Record path [–]	What was registered and not produced [–]
`classifiers`	the sixteen classifier arms (four classifiers, pooled and paired, raw and physics-residual rows)
`fingerprint_controls`	the campaign-fingerprint controls on those arms
`inflation`	the point-split inflation arm
`label_levels.one.arms`	the one-label retraining of the classifier arms
`sensitivities.descriptive_classifiers`	the gradient-boosting rows
`sensitivities.seed_spread`	the five-seed spread of the classifier arms
`sensitivities.facility_addback`	the facility-loop add-back arm
`sensitivities.` `classifier_run_mean_differencing`	run-mean differencing for the classifiers
`descriptive_comparisons`	the eight earlier descriptive comparisons against classifier arms

names the registered class of 33 of 39 scored runs, a merged accuracy of 0.846 with interval [0.695, 0.941], against 19 of 39 for the comparator, 0.487 with interval [0.324, 0.652]; the intervals do not overlap. The model's count falls to 29 without the oil instrumentation and to 26 without the refrigerant pressures, holds at 26 on the building-automation point list, and is 29 with five channels, where the flow faults are inferred from the energy balance rather than read from a meter. The comparator's count is 19 without the oil instrumentation, 17 at the two sets that leave it only its two water-flow rules, and 0 at the fifth, where it cannot run. The fold-mean offset center, a registered sensitivity, gives 31 and 28 at the first two sets.

Table 10 reads the same answers inside each refrigerant-side class, and Tables 11 and 12 name every run whose answer of (7) the model is not credited for at the first two sets, with its true family and the answer given. At the laboratory instrumentation the six uncredited runs are `normal cf5`, `cf45`, `nc trace`, `nc trace2`, `rl10`, `rl20`: the deepest fouling severity is answered as reduced condenser flow, which plugged tubes and a throttled condenser loop both resemble from the condenser's side; the two trace nitrogen runs, which have no quantified amount, and the mildest leak are answered as fouling, inside the class the companion report found inseparable; the second leak severity is answered as overcharge; and one normal run is called non-condensable gas. Without the oil instrumentation the ten uncredited runs include the three excess-oil runs, answered normal, and three normal runs of one day block, placed in the merged fouling class. The comparator's twenty uncredited runs at the laboratory instrumentation are `cf12`, `cf20`, `cf6`, `eo14`, `eo32`, `eo50`, `nc trace2`, `rl10`, `rl20`, answered normal; `cf45`, `nc trace`, `nc1`, `nc2`, `nc3`, `nc5`, `ro10`, `ro20`, `ro30`, `ro40`, answered ambiguous — overcharge and non-condensable gas share one pattern in its table, so both rules fire together; and `cf30`, answered as reduced condenser water flow.

The model and the comparator are set side by side on the full population because the question a user of the sensors asks is which method names more faults from the same readings. The gap between them is widest where the comparator cannot run at all and narrowest at the two sets that leave it only its two water-flow rules; under the protocol it is descriptive, since the confirmatory comparison runs on the fallback population of Section 4.2. The merged count is printed beside the unmerged one because merging forgives the confusions the model's own signatures predicted; at the laboratory instrumentation every family is its own class and the two counts coincide, so the model's count there owes nothing to merging. What this shows about the method is that a model fixed on normal operation and scored on runs it never saw names the fault of most scored runs from the physics alone, and keeps answering on a point list too small for the rule-based method to run.

Table 9: Run-level accuracy of every method at every sensor set on the same 39 scored runs: the merged count and the rate of (34) with its exact binomial interval, the count under the unmerged scoring (each family its own class), and the comparator's not-applicable answers, which are scored wrong. The confirmatory offset center of the model is the paired normal; the fold-mean center is a registered sensitivity. No analysis failed at any set.

Set [–]	Method [–]	Correct [–]	Accuracy [–]	95% interval [–]	Unmerged correct [–]	N/A [–]
L0	the model	33 of 39	0.846	[0.695, 0.941]	33	0
L0	the model, fold-mean center	31 of 39	0.795	[0.635, 0.907]	31	0
L0	the comparator	19 of 39	0.487	[0.324, 0.652]	19	0
L0	the comparator with the oil rule	22 of 39	0.564	[0.396, 0.722]	22	0
L1	the model	29 of 39	0.744	[0.579, 0.870]	27	0
L1	the model, fold-mean center	28 of 39	0.718	[0.551, 0.850]	27	0
L1	the comparator	19 of 39	0.487	[0.324, 0.652]	19	0
L1	the comparator with the oil rule	19 of 39	0.487	[0.324, 0.652]	19	0
L2	the model	26 of 39	0.667	[0.498, 0.809]	23	0
L2	the model, fold-mean center	26 of 39	0.667	[0.498, 0.809]	23	0
L2	the comparator	17 of 39	0.436	[0.278, 0.604]	17	0
L2	the comparator with the oil rule	17 of 39	0.436	[0.278, 0.604]	17	0
L3	the model	26 of 39	0.667	[0.498, 0.809]	22	0
L3	the model, fold-mean center	26 of 39	0.667	[0.498, 0.809]	22	0
L3	the comparator	17 of 39	0.436	[0.278, 0.604]	17	0
L3	the comparator with the oil rule	17 of 39	0.436	[0.278, 0.604]	17	0
L4	the model	29 of 39	0.744	[0.579, 0.870]	23	0
L4	the model, fold-mean center	28 of 39	0.718	[0.551, 0.850]	22	0
L4	the comparator	0 of 39	$< 10^{-9}$	$[< 10^{-9}, 0.090]$	0	39
L4	the comparator with the oil rule	0 of 39	$< 10^{-9}$	$[< 10^{-9}, 0.090]$	0	39

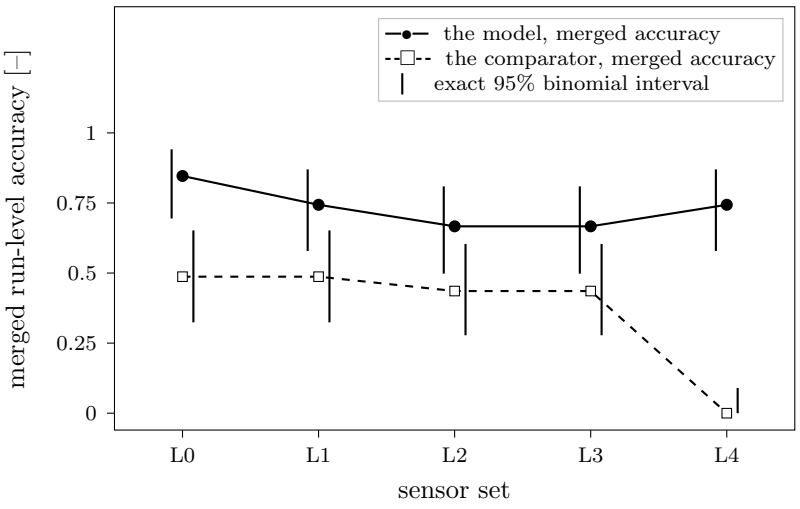


Figure 4: Merged run-level accuracy of the model (solid line, filled circles) and of the comparator (dashed line, open squares) against the sensor set, on the same 39 scored runs at every set, with exact 95 percent binomial intervals drawn as vertical bars, offset to either side of the set for legibility. Each mark is the accuracy of (34): a run is credited when the method's answer is the registered class that holds the truth at that set, and the comparator's not-applicable answers at L4 are scored wrong and counted. The model lies above the comparator at every set, by the widest margin at L4, where the comparator cannot run. The curve is descriptive and carries no test.

Table 10: Accuracy inside each registered class of the refrigerant-side families, at every sensor set, for the model and for the comparator. A multi-family class is credited when the answer is that class; the two water-flow families and normal are scored in Table 9 and are not repeated here.

Set [–]	Method [–]	Correct runs within each refrigerant-side class [–]
L0	model	`cf`: 4 of 5; `eo`: 3 of 3; `nc`: 4 of 6; `rl`: 2 of 4; `ro`: 4 of 4
L0	comparator	`cf`: 0 of 5; `eo`: 0 of 3; `nc`: 0 of 6; `rl`: 2 of 4; `ro`: 0 of 4
L1	model	{`cf`, `eo`, `rl`}: 4 of 12; `nc`: 5 of 6; `ro`: 4 of 4
L1	comparator	{`cf`, `eo`, `rl`}: 2 of 12; `nc`: 0 of 6; `ro`: 0 of 4
L2	model	{`cf`, `eo`, `rl`, `ro`}: 5 of 16; `nc`: 4 of 6
L2	comparator	{`cf`, `eo`, `rl`, `ro`}: 0 of 16; `nc`: 0 of 6
L3	model	{`cf`, `eo`, `rl`, `ro`}: 4 of 16; `nc`: 4 of 6
L3	comparator	{`cf`, `eo`, `rl`, `ro`}: 0 of 16; `nc`: 0 of 6
L4	model	{`cf`, `eo`, `rl`, `ro`}: 5 of 16; `nc`: 4 of 6
L4	comparator	{`cf`, `eo`, `rl`, `ro`}: 0 of 16; `nc`: 0 of 6

Table 11: Every scored run the model does not credit at L0 (the laboratory instrumentation), with its true family and the answer given; a multi-family answer is a merged class of that sensor set.

Run [–]	Truth [–]	The model's merged answer [–]
`normal cf5`	`normal`	`nc`
`cf45`	`cf`	`fwc`
`nc trace`	`nc`	`cf`
`nc trace2`	`nc`	`cf`
`rl10`	`rl`	`cf`
`rl20`	`rl`	`ro`

Table 12: Every scored run the model does not credit at L1 (no oil instrumentation), with its true family and the answer given; a multi-family answer is a merged class of that sensor set.

Run [–]	Truth [–]	The model's merged answer [–]
`normal cf2`	`normal`	{`cf`, `eo`, `rl`}
`normal cf4`	`normal`	{`cf`, `eo`, `rl`}
`normal cf5`	`normal`	{`cf`, `eo`, `rl`}
`cf30`	`cf`	`fwc`
`cf45`	`cf`	`fwc`
`eo14`	`eo`	`normal`
`eo32`	`eo`	`normal`
`eo50`	`eo`	`normal`
`nc trace2`	`nc`	{`cf`, `eo`, `rl`}
`rl20`	`rl`	`ro`

4.2 The confirmatory comparison

The four registered comparisons on the fallback population of 17 runs are set out in Table 13, one row per sensor set, with the four cells of the paired count, the difference Δ of (35) with its bootstrap interval, and the p-value against its Holm limit. At every set the comparator answers all 17 runs correctly — the nine water-flow faults are read by its two water-temperature-difference rules and the eight normals are not flagged — and the model answers 16 of them at L0 and 14 at each of L1 to L3, so the discordant runs are 1 and 3 in the comparator's favor and none in the model's. The differences are -0.0588 at L0 with interval [-0.143, $< 10^{-9}$] and -0.1765 at L1 to L3 with intervals whose lower ends reach -0.444; the p-values of (33) are 1.00 and 0.25, Holm rejects none of the

four, and every verdict is *not distinguishable at this sample size*. The probability of a verdict for each split of the discordant runs and each significance level, as the companion report registered it, is tabulated in Table 14; read against it, a split of three discordant runs in ten had a probability of 0.10464 of a verdict at Holm's first threshold on this population; the comparison was underpowered by construction, and its outcome is a difference of one to three runs in the comparator's favor, on a population where a flow meter reads every fault directly and a rule on one temperature difference suffices. Every discordant run is a scored normal the model calls faulty: `normal cf5` at L0, called non-condensable gas; `normal cf2`, `normal cf4`, `normal cf5` at L1, three normals of one day block placed in the merged fouling class; and `near normal1`, `normal eo`, `normal r1` at L2 and L3, placed in the merged class of the four refrigerant-side families. Every one of the nine flow-fault runs is named by both methods at all four sets.

The comparison is made on the fallback population because the protocol fixed it there once the acceptance test failed. The whole difference between its two sides lies on scored normal runs that the model calls faulty; on the faults themselves the two methods agree run for run. What this demonstrates about the method is that, where a meter reads the fault directly, the physics-structured model reaches the same fault answers as a rule built for that fault, and that its price there is false calls on normal runs, not missed faults.

Table 13: The confirmatory family: the model against the comparator at the four registered sensor sets, on the fallback population of 17 runs. *Both*, *model only*, *comparator only* and *neither* count the runs each method answered correctly; the difference Δ of (35) is the model's merged accuracy minus the comparator's, with the group-bootstrap interval; p is the exact McNemar p-value of (33) and the Holm limit the threshold it had to clear in the family of four. A claim needs the rejection, an interval on the model's side of zero and discordant runs that favor the model; every verdict is *not distinguishable at this sample size*.

Set [–]	Both [–]	Model only [–]	Comparator only [–]	Neither [–]	Δ [–]	95% interval [–]	p [–]	Holm limit [–]
L0	16	0	1	0	-0.0588	[-0.143, $< 10^{-9}$]	1.0000	0.05
L1	14	0	3	0	-0.1765	[-0.429, $< 10^{-9}$]	0.2500	0.0125
L2	14	0	3	0	-0.1765	[-0.444, $< 10^{-9}$]	0.2500	0.0166667
L3	14	0	3	0	-0.1765	[-0.444, $< 10^{-9}$]	0.2500	0.025

Table 14: What the confirmatory comparison could resolve on the fallback population, registered before any verdict was read. Two methods are compared run by run and only the runs on which exactly one of them is right carry information; for each assumed rate of runs only the model answers right and of runs only the comparator answers right, and each significance level, the last column gives the probability that the paired test of (33) rejects. Rates printed as bounds are zero to round-off.

Runs [–]	Significance level [–]	Rate, model only right [–]	Rate, comparator only right [–]	Probability of a verdict [–]
17	0.0125	0.3	$< 10^{-9}$	0.10464
17	0.0125	0.25	0.05	0.0231984
17	0.0125	0.2	$< 10^{-9}$	0.0109343
17	0.0125	0.15	0.05	0.00103106
17	0.05	0.3	$< 10^{-9}$	0.403181
17	0.05	0.25	0.05	0.129344
17	0.05	0.2	$< 10^{-9}$	0.105701
17	0.05	0.15	0.05	0.0172591

4.3 Detection and false calls

Detection by family and false calls, the rates of (36) and (37), are counted for both methods at their registered operating points in Table 15, one row per method and set; the model's false calls are compared with the rates predicted in advance in Table 16, with exact intervals and the runs called. With the laboratory instrumentation the model answers *not normal* on 31 of 31 fault runs, every family at every severity, with 1 false call in 8 scored normals and 3 in 17; the comparator detects 22 of 31 with no false call, missing `cf6`, `cf12`, `cf20`, `rl10`, `rl20`, `eo14`, `eo32`, `eo50`, `nc trace2`: the three mildest fouling severities, every excess-oil run, which it has no rule for, the two mildest leaks and one trace nitrogen run. At the lower sets the model's detection falls to 23 of 31 without the refrigerant pressures and recovers to 25 with five channels, while its false calls rise to 3 of 8 scored normals from the second set on. The predicted false-call rate was 0.074 with every sensor against a realized 0.125 on the scored normals with interval [0.003, 0.527], and 0.41 to 0.39 at the three lower sets against a realized 0.375 with interval [0.085, 0.755]: each prediction lies inside the realized interval. The comparator makes no false call at any set, at the price of the detections it misses. Table 17 stratifies the model's accuracy by the pairing gap and the day block; re-pairing the scored normals within their own block changes 0 answer at L0 and 2 at the sets without refrigerant pressures.

Detection and false calls are compared between the two methods because each method is set at its own registered operating point and the two rates trade against each other. The model detects more fault runs at every set where both run and calls more normal runs faulty; the comparator calls none faulty and misses the mildest severities and every excess-oil run. The false calls are compared with the rates predicted in advance to test whether the model's behavior on an untroubled machine was foreseeable from normal operation alone; that every prediction lies inside the realized interval shows the false-call rate is a property of the method that can be planned for before the first fault occurs.

4.4 Magnitude intervals

Figure 5 plots the model's 95 percent magnitude interval for each of the nine water-flow fault runs against the magnitude its nominal severity implies, at the laboratory instrumentation and at the set without flow meters, one vertical bar per run with the diagonal marking an interval value equal to the nominal; the coverage κ of (38) at every set, with its exact interval and the rule's verdict, is in Table 18. With the meters read, every interval is narrow — its half-width is at most 6.9 percent of the nominal magnitude — and seven of the nine lie below the diagonal, by 0.6 to 12.1 percent of the nominal: 2 of 9 intervals cover at L0 and 3 of 9 at L1, so the registered rule fails at both sets and the magnitude claim is not made. At the set without flow meters the intervals widen, because the flow is then inferred from the energy balance, and 6 of 9 cover. The other five families, reported beneath each set in the table, cover poorly at the laboratory instrumentation — fouling and excess oil in part, the charge faults and the gas not at all — and better at the lower sets, where the intervals are wider; their maps depend on structural constants, and their coverage tests those constants as much as the inference.

The intervals are compared with the nominal severities because those are the only magnitudes the benchmark records for its fault runs. The difference between the two is a shortfall shared by most runs, narrow intervals sitting low together, rather than scatter about the diagonal. What this demonstrates is that the model's magnitude is precise, and that whether it is also accurate turns on a reference this benchmark does not supply; the two readings of the shortfall are taken up in Section 5.

Table 15: Detection — any answer but normal on a fault run, the rate d of (36) — and false calls — any answer but normal on a normal run, the rate f of (37) — for the model and the comparator, each at its registered operating point. The model answers with its most probable hypothesis under equal prior odds; the comparator flags a point when any feature's standardized residual exceeds its threshold, set for a five percent false-alarm rate on held-out normal points. False calls are counted on the eight scored normals and on all seventeen usable normals, each paired outside its own day block. Descriptive; not in the confirmatory family.

Set [–]	Method [–]	Detected [–]	By family [–]	False calls, 8 scored [–]	False calls, all 17 [–]
L0	model	31 of 31	`cf` 5/5; `fwc` 5/5; `fwe` 4/4; `rl` 4/4; `ro` 4/4; `eo` 3/3; `nc` 6/6	1 of 8	3 of 17
L0	comparator	22 of 31	`cf` 2/5; `fwc` 5/5; `fwe` 4/4; `rl` 2/4; `ro` 4/4; `eo` 0/3; `nc` 5/6	0 of 8	0 of 17
L1	model	28 of 31	`cf` 5/5; `fwc` 5/5; `fwe` 4/4; `rl` 4/4; `ro` 4/4; `eo` 0/3; `nc` 6/6	3 of 8	6 of 17
L1	comparator	22 of 31	`cf` 2/5; `fwc` 5/5; `fwe` 4/4; `rl` 2/4; `ro` 4/4; `eo` 0/3; `nc` 5/6	0 of 8	0 of 17
L2	model	23 of 31	`cf` 4/5; `fwc` 5/5; `fwe` 4/4; `rl` 2/4; `ro` 4/4; `eo` 0/3; `nc` 4/6	3 of 8	6 of 17
L2	comparator	16 of 31	`cf` 3/5; `fwc` 5/5; `fwe` 4/4; `rl` 0/4; `ro` 2/4; `eo` 0/3; `nc` 2/6	0 of 8	0 of 17
L3	model	23 of 31	`cf` 4/5; `fwc` 5/5; `fwe` 4/4; `rl` 2/4; `ro` 4/4; `eo` 0/3; `nc` 4/6	3 of 8	6 of 17
L3	comparator	16 of 31	`cf` 3/5; `fwc` 5/5; `fwe` 4/4; `rl` 0/4; `ro` 2/4; `eo` 0/3; `nc` 2/6	0 of 8	0 of 17
L4	model	25 of 31	`cf` 4/5; `fwc` 5/5; `fwe` 4/4; `rl` 3/4; `ro` 4/4; `eo` 1/3; `nc` 4/6	3 of 8	6 of 17
L4	comparator	not applicable	—	—	—

Table 16: The model's false calls on normal runs, the rate f of (37), against the rate the companion report predicted for each sensor set from simulated normal operation before any fault run was read. Counts are over the eight scored normals and over all seventeen usable normals, with exact binomial intervals; the last column names the runs called faulty among the seventeen.

Set [–]	Predicted rate [–]	Scored [–]	95% interval [–]	All usable [–]	95% interval [–]	Runs called [–]
L0	0.074	1 of 8	[0.003, 0.527]	3 of 17	[0.038, 0.434]	`normal cf5`, `normal cf6`, `normal r`
L1	0.149	3 of 8	[0.085, 0.755]	6 of 17	[0.142, 0.617]	`normal b`, `normal cf2`, `normal cf4`, `normal cf5`, `normal cf6`, `normal r`
L2	0.415	3 of 8	[0.085, 0.755]	6 of 17	[0.142, 0.617]	`near normal1`, `near normal3`, `normal cf`, `normal eo`, `normal r1`, `normal2`
L3	0.414	3 of 8	[0.085, 0.755]	6 of 17	[0.142, 0.617]	`near normal1`, `near normal3`, `normal cf`, `normal eo`, `normal r1`, `normal2`
L4	0.392	3 of 8	[0.085, 0.755]	6 of 17	[0.142, 0.617]	`near normal1`, `near normal3`, `normal cf`, `normal eo`, `normal r1`, `normal2`

Table 17: The model's merged accuracy of (34), as correct counts, stratified by the pairing gap — the days between a scored run and the normal run it is paired with — and by whether the pair lies in one day block. The last column counts the scored normals whose answer changes when they are re-paired with a partner from their own block, a registered sensitivity of the offset center.

Set [–]	Gap at most 1 d [–]	Gap of 2–7 d [–]	Gap over 7 d [–]	Same block [–]	Other block [–]	Changed by re-pairing [–]
L0	6 of 7	22 of 24	5 of 8	6 of 7	27 of 32	0
L1	4 of 7	18 of 24	5 of 8	4 of 7	23 of 32	1
L2	4 of 7	13 of 24	6 of 8	4 of 7	19 of 32	2
L3	4 of 7	12 of 24	6 of 8	4 of 7	18 of 32	2
L4	5 of 7	12 of 24	6 of 8	5 of 7	18 of 32	2

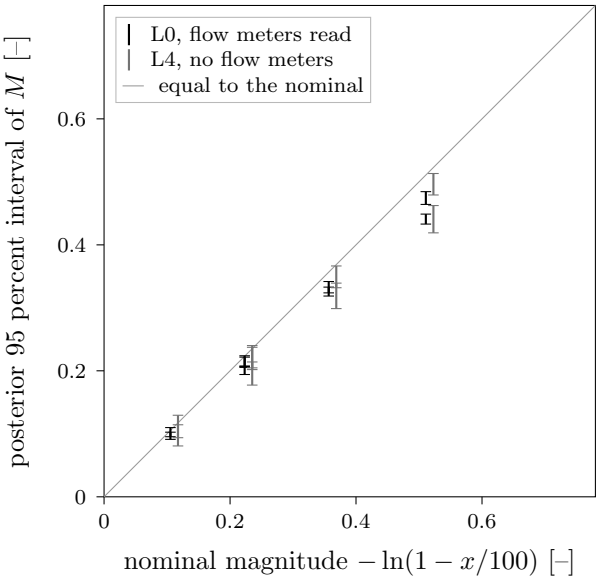


Figure 5: The model's 95 percent magnitude interval for each water-flow fault run, under the run's true hypothesis, against the magnitude the run's nominal severity implies, $M = -\ln(1 - x/100)$ for a flow reduced by x percent. Black bars are the laboratory instrumentation, where the flow meter is read; gray bars, offset to the right, are the set without flow meters, where the flow is inferred from the energy balance. An interval covers in the sense of (38) when it crosses the gray diagonal. With the meters read the bars are short and most sit just below the diagonal; without them they are longer and more of them reach it.

4.5 Calibration

The calibration rule, the mean stated confidence of (39) against the top-class accuracy of (40), is evaluated at every set in Table 19, beside the Brier and log scores of (41) and (42). The reliability diagram at the two registered sets, each bin's mean confidence against its accuracy, is shown in Figure 6, and its bins are listed in Table 20. At the laboratory instrumentation the mean stated confidence of the most probable merged class is 0.910 against an accuracy of 0.846 with interval [0.695, 0.941], and without the oil instrumentation 0.847 against 0.718 with interval [0.551, 0.850]; the rule passes at both, and the calibration claim holds. The rule also passes at the three lower sets, where it is not claimed. The reliability bins show where the confidence sits: at L0, 27 of 39 runs are answered with confidence above 0.95 and 93 percent of them are right, and at L1 23 runs with 87 percent right; the middle bins hold few runs and sit at or below the diagonal. The Brier score rises from 0.264 at L0 to 0.466 without the refrigerant pressures and the log score from 1.34

Table 18: Coverage of the model's 95 percent magnitude intervals by the nominal magnitude, the fraction κ of (38). The headline families are the two water-flow faults, whose map from severity to magnitude is exact; the rule passes at a set when the exact binomial interval of their coverage includes 0.95, and is claimed when it passes at both registered sets. The other families, whose map depends on a structural constant, are reported beneath each set and govern nothing.

Set [–]	Families [–]	Covered [–]	Coverage [–]	95% interval [–]	Passes [–]	Registered [–]
L0	`fwc`, `fwe`	2 of 9	0.222	[0.028, 0.600]	no	yes
	other families	`cf` 2/5; `eo` 1/3; `nc` 0/4; `rl` 0/4; `ro` 0/4				
L1	`fwc`, `fwe`	3 of 9	0.333	[0.075, 0.701]	no	yes
	other families	`cf` 2/5; `eo` 3/3; `nc` 0/4; `rl` 0/4; `ro` 1/4				
L2	`fwc`, `fwe`	3 of 9	0.333	[0.075, 0.701]	no	no
	other families	`cf` 4/5; `eo` 3/3; `nc` 0/4; `rl` 2/4; `ro` 2/4				
L3	`fwc`, `fwe`	3 of 9	0.333	[0.075, 0.701]	no	no
	other families	`cf` 4/5; `eo` 3/3; `nc` 0/4; `rl` 2/4; `ro` 2/4				
L4	`fwc`, `fwe`	6 of 9	0.667	[0.299, 0.925]	no	no
	other families	`cf` 4/5; `eo` 3/3; `nc` 0/4; `rl` 2/4; `ro` 2/4				

to 2.29; no truth probability underflowed to zero at any set. Pooling fouling and non-condensable gas into one class at L0 lowers the Brier score to 0.187.

The stated confidence is compared with the realized accuracy because a diagnosis is acted on through its probability: a calibrated posterior lets a user set a threshold that trades false calls for missed faults. The two agree within the binomial interval at every set, and the mean stated confidence falls from the laboratory instrumentation to the lower sets as information is removed. This demonstrates that the evidence comparison of (6) yields probabilities that can be read as frequencies on runs the model never saw.

Table 19: Calibration of the model's posterior on the merged classes at every sensor set. The rule passes when the mean stated confidence $\bar{\pi}$ of the most probable class, (39), lies inside the exact binomial interval of that class's accuracy a_{top}, (40); it is claimed when it passes at L0 and L1. The Brier score B of (41) is the mean squared distance between the posterior and the truth indicator; the log score S of (42) is the mean of $-\ln$ of the probability given to the truth, printed with its median and the count of runs whose truth probability underflowed to zero.

Set [–]	Top class right [–]	Accuracy, 95% interval [–]	Mean confidence [–]	Passes [–]	Brier [–]	Log score [nats]	Median [nats]	Infinite [–]
L0	33 of 39	[0.695, 0.941]	0.910	yes	0.264	1.339	$< 10^{-9}$	0
L1	28 of 39	[0.551, 0.850]	0.847	yes	0.366	1.486	0.050	0
L2	27 of 39	[0.524, 0.830]	0.806	yes	0.466	2.290	0.334	0
L3	27 of 39	[0.524, 0.830]	0.807	yes	0.467	2.314	0.336	0
L4	29 of 39	[0.579, 0.870]	0.801	yes	0.348	0.555	0.335	0

4.6 The comparator in detail

Table 21 sums the comparator's point-level outcomes in its author's categories, one column per category, and its merged run-level count of (34) for every registered variant is in Table 22. At the laboratory instrumentation its threshold c lies between 3.11 and 3.40 in the standardized residuals of (32) across the folds, and its held-out false-alarm rate on normal points is within a tenth of a percent of the registered five percent in every fold, so the reproduction behaves as the method is described. Correct-but-not-unique outcomes are numerous there because overcharge and non-condensable gas share one five-feature pattern, as the author states; a run where both rules fire

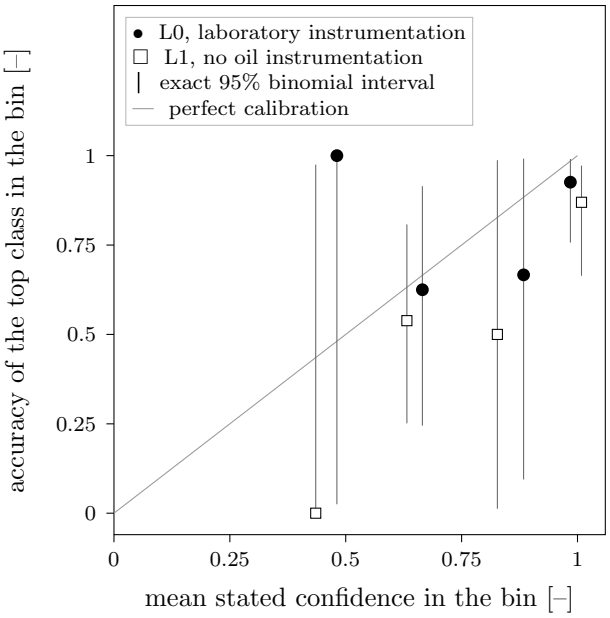


Figure 6: Reliability of the model's stated confidence at the two sensor sets where the calibration claim is registered. Runs are binned by the probability the model gives its most probable merged class; each point is one bin's mean stated confidence against the fraction of its runs the top class was right on, the two quantities compared by the calibration rule of (39), with an exact 95 percent binomial interval. The two sets' points are offset slightly to either side of the bin's confidence so that both stay visible; the bins' run counts are listed in Table 20. A calibrated model lies on the gray diagonal and a point below it is overconfident.

Table 20: The reliability bins of Figure 6: at L0 and L1, runs binned by the probability of their most probable merged class, each bin's mean stated confidence against the fraction of its runs the top class was right on, with an exact binomial interval. An empty bin prints no accuracy.

Set [–]	Confidence bin [–]	Runs [–]	Mean confidence [–]	Accuracy [–]	95% interval [–]
L0	[0.125, 0.5]	1	0.493	1.000	[0.025, 1.000]
L0	[0.5, 0.8]	8	0.677	0.625	[0.245, 0.915]
L0	[0.8, 0.95]	3	0.896	0.667	[0.094, 0.992]
L0	[0.95, 1]	27	0.997	0.926	[0.757, 0.991]
L1	[0.125, 0.5]	1	0.423	$< 10^{-9}$	[$< 10^{-9}$, 0.975]
L1	[0.5, 0.8]	13	0.620	0.538	[0.251, 0.808]
L1	[0.8, 0.95]	2	0.815	0.500	[0.013, 0.987]
L1	[0.95, 1]	23	0.997	0.870	[0.664, 0.972]

is ambiguous and scored wrong under the registered classes. The excess-oil extension raises the comparator's L0 count from 19 to 22 and nothing elsewhere; the alternative false-alarm rates and the relaxed match move the count by at most one; and the alternative reading that places the evaporator-flow symptom on the condenser water difference drops it to 10, because the two flow faults then alias. The conductance screen set aside one training steady state at L0 and L1 and one point of two runs; the thresholds and answers above are computed without them.

The comparator is examined on its own terms because a weak reproduction would flatter the model. Its threshold holds the registered false-alarm rate in every fold, and the alternative false-alarm rates and the relaxed symptom match move its count by at most one run. This shows that the model is compared with a faithful reproduction of the method, so that the differences of Sections 4.1

and 4.2 are differences between the two methods rather than artifacts of the reproduction.

Table 21: The comparator's point-level outcomes in its author's own categories, summed over every scored run's steady states at each sensor set. A detected point with exactly one rule fired on the true fault is correct and unique; with several rules fired, the true one among them, correct but not unique; with none fired, unable to diagnose; with rules fired and the true one absent, incorrect. The last three columns are the undetected fault points and the two outcomes on normal points.

Set [–]	Correct, unique [–]	Correct, not unique [–]	Unable to diagnose [–]	Incorrect [–]	Not detected [–]	False alarm [–]	True negative [–]
L0	278	171	91	42	239	22	433
L1	278	171	91	42	239	22	433
L2	235	5	8	166	408	18	438
L3	238	5	8	177	409	18	440
L4	not applicable: the load regressor is removed						

Table 22: Merged run-level correct counts, of 39, for the comparator as registered, for the excess-oil extension and for each registered sensitivity — the two alternative false-alarm rates, the relaxed symptom match and the three alternative readings of the table cells the accessible text leaves ambiguous — at every sensor set. Only the first row is in the confirmatory family.

Variant [–]	L0 [–]	L1 [–]	L2 [–]	L3 [–]	L4 [–]
the comparator as registered	19	19	17	17	0
with the excess-oil rule added	22	19	17	17	0
false-alarm rate 0.025	18	18	17	17	0
false-alarm rate 0.1	19	19	17	17	0
two of three symptoms suffice	18	18	17	17	0
fouling read without the approach cell	19	19	17	17	0
fouling read without the conductance cell	19	19	17	17	0
evaporator flow read on the condenser water dT	10	10	8	8	0

4.7 Hard cases

The model's answers of (7) on the four hard cases are listed in Table 23, one row per run and one column per sensor set. The mildest fouling severity is named at the laboratory instrumentation (`cf`) and answered inside the merged class or as overcharge at the lower sets; the mildest evaporator-flow fault is named at every set; the mildest leak is answered as fouling where the refrigerant channels are present and as normal or overcharge where they are not; and the trace nitrogen runs, whose amount the benchmark did not record, are answered as fouling at L0, as nitrogen at L1 and as normal at the three lower sets. Every one of these lies inside the identifiability class the companion report computed for its severity, or inside the scatter it marked, and none counts for or against any claim.

The hard cases are compared with the companion report's identifiability classes because those classes state in advance which confusions the instrumentation forces. That every answer falls inside them, or inside the scatter the companion report marked, shows that the model's errors on the mildest faults are the ones its own physics predicted, which is what lets a user know beforehand which answers to read as a class rather than as a single fault.

Table 23: The four hard cases — the mildest quantified severity of four families — with the model's merged answer at every sensor set. They are scored in the population and are never averaged into a headline on their own.

Run [–]	Truth [–]	L0 [–]	L1 [–]	L2 [–]	L3 [–]	L4 [–]
cf6	cf	cf	eo	ro	ro	ro
fwe10	fwe	fwe	fwe	fwe	fwe	fwe
rl10	rl	cf	cf	normal	normal	ro
nc trace	nc	cf	nc	normal	normal	normal

4.8 The detection predictions against the outcomes

The companion report's detection predictions are set against the outcomes in Table 24: for each sensor set, the range ends checked and those that disagree, and the count of detected runs against the expectation E of (43) with its p-value. No run's outcome had probability below the registered level under its prediction at any set, and the count tests agree: at the laboratory instrumentation 19 of 19 predictable runs were detected against an expectation of 18.4, and with five channels 25 of 29 against 22.1, with two-sided p-values of 1.00 and 0.19. The disagreements at the range ends all run one way: a family's mildest severity marked as expected undetectable — fouling and the gas at every set, the charge faults at the lower sets — was detected, 2 such ends at L0 and 6 with five channels, against one expected-detectable end not detected, the deepest excess-oil run without the oil sensor. A mild fault detected where the prediction said it would sit inside the scatter is a prediction that was too cautious, not a false claim; it is reported as a disagreement.

The outcomes are compared with predictions made before any fault run was read because a prediction that could have failed and did not is evidence that the model's account of its own sensitivity is right. That every disagreement runs toward more detection than predicted shows the predictions erred on the cautious side, the direction that does not overstate to a user what the instrumentation can see.

4.9 Context arms and sensitivities

The label-free context arms are compared with the model in Table 25: detections and false calls for the two detectors, and correct counts in the sense of (34) for the two arms that name a family. At the laboratory instrumentation the two detectors on the model's residual rows mark 20 and 19 of 31 fault runs with no false call, against the model's 31; the directional arm names 29 of 39 runs correctly and the sign rule table 21, against the model's 33. The gap narrows at the lower sets, where every arm reads the same few channels. Table 26 counts the runs whose answer changes under each registered physics sensitivity: at most 6 of 48 at any set, most often under the doubled structural prior widths at the lower sets and the shifted design approaches where the refrigerant pressures are absent. No sensitivity changes a confirmatory verdict, which are computed from the registered settings only.

The context arms are set beside the model because they read the same residual rows without weighing whole hypotheses against one another; the gap between them and the model at the laboratory instrumentation measures what the evidence comparison of (6) adds to the same residuals. The sensitivities are counted to show which answers rest on a setting registered with a range rather than on the readings themselves.

Table 24: The detection predictions of the companion report against the outcomes. Each family's lowest and highest quantified severity was predicted detectable or not at each sensor set; the table counts the range ends checked, after setting aside the outcomes known in advance, and those whose outcome disagrees with the prediction. The count test compares the number of runs detected with the sum E of (43) of the per-run detection probabilities the companion report simulated, with the Poisson-binomial two-sided p-value. No per-run contradiction at the registered level of 0.05 occurred at any set.

Set [–]	Range ends [–]	Disagree [–]	Detected [–]	Expected [–]	p [–]	Disagreements at the range ends [–]
L0	9	2	19 of 19	18.40	1.000	`cf` at lowest: predicted undetectable, detected; `nc` at lowest: predicted undetectable, detected
L1	10	3	17 of 20	17.30	1.000	`cf` at lowest: predicted undetectable, detected; `eo` at highest: predicted detectable, not detected; `nc` at lowest: predicted undetectable, detected
L2	10	4	14 of 20	13.45	1.000	`cf` at lowest: predicted undetectable, detected; `rl` at highest: predicted undetectable, detected; `ro` at lowest: predicted undetectable, detected; `nc` at lowest: predicted undetectable, detected
L3	10	4	14 of 20	13.45	1.000	`cf` at lowest: predicted undetectable, detected; `rl` at highest: predicted undetectable, detected; `ro` at lowest: predicted undetectable, detected; `nc` at lowest: predicted undetectable, detected
L4	14	6	25 of 29	22.13	0.186	`cf` at lowest: predicted undetectable, detected; `rl` at lowest: predicted undetectable, detected; `rl` at highest: predicted undetectable, detected; `ro` at lowest: predicted undetectable, detected; `eo` at highest: predicted undetectable, detected; `nc` at lowest: predicted undetectable, detected

Table 25: The label-free context arms on the model's own residual rows, at every sensor set: two detectors — Hotelling's T^2 bound and a one-class support-vector machine, each at a five percent false-alarm rate — with their detections on the 31 fault runs and their false calls on the eight scored normals; the directional arm, which names the family whose expected signature is nearest the run's mean residual; and the sign rule table on the same residuals. The model's merged count is repeated for comparison. Descriptive only; the directional arm's answer on a few nitrogen runs at the lower sets depends on the linear-algebra kernel (Appendix A).

Set [–]	T^2 detects [–]	T^2 false [–]	SVM detects [–]	SVM false [–]	Directional correct [–]	Rule table correct [–]	Model correct [–]
L0	20 of 31	0 of 8	19 of 31	0 of 8	29 of 39	21 of 39	33 of 39
L1	18 of 31	0 of 8	17 of 31	0 of 8	26 of 39	18 of 39	29 of 39
L2	16 of 31	0 of 8	14 of 31	0 of 8	18 of 39	16 of 39	26 of 39
L3	16 of 31	0 of 8	15 of 31	0 of 8	18 of 39	16 of 39	26 of 39
L4	14 of 31	0 of 8	14 of 31	0 of 8	18 of 39	17 of 39	29 of 39

Table 26: The registered physics sensitivities of the model: each is the confirmatory computation with one structural input or prior moved to the end of its registered range, and the cells count the 48 diagnosed runs (31 fault runs and 17 normals) whose most probable hypothesis changed at each sensor set; a dash marks an arm the protocol does not run at that set. None governs any claim.

Sensitivity [–]	Variant [–]	L0 [–]	L1 [–]	L2 [–]	L3 [–]	L4 [–]
charge inventory at the ends of its registered ranges	overfeed share top	3	2	3	2	2
charge inventory at the ends of its registered ranges	superheat authority full	0	0	0	1	0
condensing-pressure drift prior at the ends of its span	0	2	1	—	—	—
condensing-pressure drift prior at the ends of its span	2	0	1	—	—	—
conductance drift at the supported covariance	—	2	5	0	0	0
non-condensable gas moving the conductance instead of the pressure	—	4	2	—	—	—
structural prior widths at half and at twice their values	0.5	1	0	0	1	0
structural prior widths at half and at twice their values	2	0	2	3	4	6
design approaches shifted by one log-sd	-1	—	—	1	0	1
design approaches shifted by one log-sd	1	—	—	2	3	5
design flows at ±10 percent	0.9	—	—	—	—	0
design flows at ±10 percent	1.1	—	—	—	—	2

5 Discussion

The central result is that a model fixed on normal operation alone, shown no fault example, names the registered class of 33 of 39 runs with the laboratory instrumentation and 29 with five channels, with a stated confidence that matches its accuracy at the sets where that was registered, and that it does so with one false call in eight normal runs at the full instrumentation and three in eight below. The registered comparison does not establish that it outperforms the rule-based comparator, and it could not have: the fallback the failed acceptance test triggered confines the comparison to the nine runs where a flow meter reads the fault directly, on which a rule on one water temperature difference is complete, and the model is one to three runs behind it there. The full-population counts of Table 9 and Figure 4, whose intervals do not overlap at the first sensor set or at the fifth, are descriptive under the protocol, and this report words them as such.

The failed magnitude rule deserves the most care, because it is a precise failure. The model's intervals for the flow faults are narrow, which is what reading the fault's own flow meter should give, and seven of nine sit 1 to 12 percent below the magnitude the nominal severity implies. The nominal severity is a number in a file name; the magnitude the model reports is the metered flow against the training normals' metered flow. A systematic shortfall of this size says either that the throttled flow in the benchmark differed from its nominal by that much, or that the training normals' flow is not the fault run's unthrottled flow, or both; the reference the rule was registered against is the file name, and the rule fails against it. At the set without flow meters, where the flow is inferred from the energy balance, the intervals widen and six of nine cover. What this report cannot say, because the protocol did not register it, is which of the two readings of the shortfall is right; the metered flows of the fault runs are in the published data and a later analysis can settle it.

Three disclosures bear on the comparison. The comparator carries design-time label information from the very runs it is scored on, a leak that favors it; the merged classes come from the model's own signatures, which removes the model's structural confusions by construction and is why the unmerged counts are printed beside every merged one; and the comparator as first built stopped with an error on a published row, which a run of the comparator on every published row before registration would have found. The repair was made after the normal-run stage, with nothing of the failed fault-run computation read beyond its failure list, and the amendment says so; a reader who weighs that disclosure differently has the record of Table 7 to weigh it with. The dropped classifier arms leave one registered question unanswered here, whether a label-trained method on these runs recognizes the campaign rather than the fault; the leakage control that would have answered it is not reported, and the literature's accuracies are cited as context only.

The model's failure modes are specific. Where the refrigerant pressures are present it confuses the deepest fouling with a condenser-flow fault, the mildest leak and the trace nitrogen runs with fouling, and the second leak severity with overcharge, each a confusion the companion report's identifiability analysis anticipated or its structural constants allow. Where the oil sensor is absent the excess-oil runs are answered normal, because the oil's only remaining path to the readings is a coupling into the evaporator whose prior puts half its mass at zero; and where the refrigerant pressures are absent, the false-call rate on normals rises to the level predicted in advance, since several fault hypotheses then explain an untroubled run about as well as normal does under equal prior odds. An operator would not run the flat rule this report scores; a decision threshold or a prior weighted by fault frequency would trade those false calls for missed mild faults, and the calibrated posterior is what makes that trade possible.

The limits of this report are plain. It concerns one machine in a laboratory, 31 fault runs and 17 normal runs; most families have four or five runs, so every per-family count carries a wide

interval. The confirmatory population is seventeen runs and the comparison on it is underpowered by construction. The magnitude reference is nominal. And the model's weaknesses on normal operation, which the companion report found and this report inherits, were not repaired between the two: the protocol froze the model, and this report scores the model that was frozen.

The answers to the five questions of this report are collected in Table 27, one row per question in the order of the Conclusion: the answer in one word with its qualifier, the number that decides it, and the measurement or analysis that would change it.

Table 27: The answers of this report, one row per question in the order of the Conclusion: the answer (yes, conditional, no or confounded) with its qualifier, the number that decides it and the equation that computed it, and the measurement or analysis that would change it. Every number is the one stated in Section 4.

Question [–]	Answer [–]	Why [–]	What would change it [–]
Does the model name the fault without fault labels?	*yes*, at the registered classes	33 of 39 scored runs at L0 and 29 of 39 with five channels, by (34)	fault runs of other machines; this record holds one
Is it more accurate than the comparator on the registered population?	*no*, not distinguishable at this sample size	$\Delta = -0.0588$ at L0 on 17 runs, by (35); Holm rejects none of four	the full population of 39 runs, open to the comparison once a refrigerant-side model passes its acceptance test
Can its magnitude intervals be trusted?	*no*, against the nominal severities	2 of 9 cover at L0 and 3 of 9 at L1, by (38); seven of nine sit low at L0	a magnitude reference read from the benchmark's metered flows
Can its stated confidence be trusted?	*yes*, at L0 and L1	mean confidence 0.910 inside [0.695, 0.941] at L0, by (39) and (40)	nothing identified
Do the detection outcomes agree with the predictions?	*yes*, at every set	no run below the level 0.05; 19 of 19 detected against 18.4 expected at L0, by (43)	nothing identified

6 Conclusion

This report asked, of a physics-structured Bayesian diagnosis model fixed on 17 normal runs of a laboratory chiller and never shown a fault, five questions about the benchmark's 31 fault runs and eight scored normal runs, with the criteria fixed before any of them was read and with the registered fallback of a failed acceptance test in force.

Whether it names the fault without labels: it names the registered class of 33 of 39 runs with the laboratory instrumentation, 29 without the oil instrumentation, 26 without the refrigerant pressures and 29 with five channels, with one false call in eight normals at the first set and three below.

How it compares with a rule-based method from the benchmark's literature: on the registered population of seventeen runs the comparator is right on every run and the model on fourteen to sixteen, a difference not distinguishable from zero at any of the four sets; on all 39 runs the model's count is 33 against 19, and the comparator cannot run with five channels, where the model's count is 29.

Whether its magnitude intervals can be trusted: not against the nominal severities, at the strictness set in advance. The flow-fault intervals are precise to within 7 percent, and seven of nine sit 1 to 12 percent low, so the coverage rule fails at both registered sets.

Whether its stated confidence can be trusted: yes, at both registered sets, where the mean confidence lies inside the interval of the accuracy; and at the three lower sets besides.

Whether the detection outcomes agree with what was predicted: no run contradicted its prediction at any set, the detected counts agree with their expectations, and every disagreement at a severity range end is a mild fault detected where it had been predicted to sit inside the scatter.

The next step is fixed by what failed: a model that represents the day-to-day shift of the refrigerant side the acceptance test caught, registered again and run on the full population, and a magnitude reference read from the benchmark's metered flows rather than from its file names.

A Reproduction

The computation on the fault runs is one command, run once,

```
python run.py stage2 --jobs 5
```

under the environment settings `OPENBLAS_CORETYPE=Haswell`, `OPENBLAS_NUM_THREADS=1`, in which the five sensor sets run as parallel processes and are merged byte for byte into the serial result. It was run through a driver script in the study directory that runs the same per-set computation in one process per set and writes each set's part as it finishes, so that a stop costs one set rather than all; it took 1.01375 hours of wall time on the machine described below, the longest set taking 60.4547 minutes. The registered timing is the micro-benchmark — one fold's confirmatory pass of the model and one classifier arm, repeated 5 times with identical outputs — whose minimum is 83.5009 seconds in a range of 83.5009 to 84.4553. The report is regenerated from the results file by a second command and contains no hand-entered number. The bootstrap uses seed 7 with 10,000 draws; the model is deterministic. The software was Python 3.12.4 with NumPy 2.4.6, SciPy 1.15.2 and scikit-learn 1.5.1 on Windows-11-10.0.26200-SP0.

Quantities at the round-off floor are printed as bounds rather than as digits, because their digits depend on which linear-algebra kernel the machine selects. To show that no *verdict* depends on that choice, the whole computation was repeated under a second kernel (`OPENBLAS_CORETYPE=SkylakeX`, `OPENBLAS_NUM_THREADS=1`) and the two runs' verdicts compared: of 435 verdicts — every method's merged answer on every scored run at every set, the four confirmatory verdicts and Holm rejections, every pass rule at every set and the two claims, every hard-case answer and the count of contradictions at every set — zero differ, and the largest move in any confirmatory p-value is 0 at the six significant digits both runs publish. The full regeneration check between the two results files — every count, label and verdict exactly, every floating-point value to a relative 1e-6 — finds 12 exact differences, all of them in the label-free directional context arm of Table 25: its answer on six nitrogen runs at L2, L3 and L4, where two candidate signatures lie at nearly the same angle to the run's residual, and the flags that record them. Those answers are kernel-dependent and are reported as such. It also finds 831 floating-point values beyond the tolerance, in the per-hypothesis magnitude details of the model's diagnoses, in the posteriors of the sensitivity arms and in two calibration scores; no merged answer, verdict, pass rule or published six-digit headline number among them.

B Registration record

The protocol file was committed on 2026-09-22; the normal-run results embed its content hash as of their own commit, and the fault-run results embed the hash of the amended file,

```
ea6668f5c2caed5d4927b8eb608013b8
9555cb975443079941e2e24bc9d55f3c
```

over line-feed-normalized bytes, together with the hash of the normal-run results they read back. Table 7 carries the commits; the computation refuses to start unless the protocol is committed and the model, the comparator and the stages are unmodified, and it records both. The protocol file itself lists everything that had been computed on normal operation before it was fixed, what each of its three amendments changed and did not change, and, for the two amendments after the normal-run stage, exactly what had been seen when each was written.

C Data and open problems

The measurements are the committed extract of the ASHRAE RP-1043 workbooks (Comstock and Braun 1999), carried with a checksum manifest and read only through it. Four problems are left open by this report. The shortfall of the flow-fault magnitudes against their nominal severities has two readings, and the metered flows that would decide between them were not registered as a reference. The classifier arms and their leakage control were dropped, so whether a label-trained method on these runs recognizes the campaign rather than the fault is not answered here. The comparator's conductance screen is a repair made after the protocol; a reproduction that keeps the original behavior will fail on the published data exactly as the first computation did. And the two published steady states the screen sets aside are physically impossible as recorded, which the benchmark's documentation does not explain.

D Notation

The symbols of the report, with their meanings, are collected in Table 28.

Table 28: The symbols of the report and their meanings.

Symbol [–]	Meaning [–]
h	a hypothesis: normal, or one of the seven fault families
M, $u = \ln M$	a fault's magnitude and its log, one per fault hypothesis
z, $r_k(z)$, w_k	the state vector, the k-th residual and its weight (inverse variance)
$C(z)$, z^*	the weighted sum of squared residuals and its minimizer, the MAP state
J, W, Λ	the Jacobian of the residuals, the weight matrix and the information matrix
m_h, s_h	the prior mean and standard deviation of u under hypothesis h
Z_h, $Z_h(\xi)$	the evidence of hypothesis h, and the same with a structural constant fixed at ξ
ξ, $p(\xi)$	the gas's segregated share or the oil coupling, integrated out of the evidence by quadrature, and its prior density
ψ, f_p, Γ_ψ	the condenser water-film share, the fraction of tubes plugged, and the factor the share's range widens the fouling prior by
$P(h \mid y)$, $\hat{h}$	the posterior probability of hypothesis h given a run's readings y, and the run's answer
N, τ_r, $\hat{H}_r$	the number of scored runs, run r's true family and its answer as a merged class
$\mathbf{1}[\cdot]$	the indicator: one when the statement in brackets holds, zero otherwise
$\hat{h}_r$	a method's answer on run r
K, K_r, $P_r(K)$	a merged class, the class holding run r's truth, and the model's probability of K
$\tilde{H}_r$, π_r	the merged class the model makes most probable for run r, and its probability
n_F, n_0, n	the numbers of fault runs, normal runs and runs counted by a measure
x_r, M_r, $[l_r, u_r]$	a flow fault's nominal reduction in percent, its nominal magnitude, and the 95 percent interval
A, a_{top}, $\bar{\pi}$	merged accuracy, top-class accuracy and mean stated confidence
d, f, κ	detection rate, false-call rate and magnitude coverage
B, S	the Brier score and the log score
p_r, E	a run's predicted detection probability and the expected detected count
b_c, ν	a run's offset on channel c and its nuisance vector
I, $\tilde{y}_{c,i}$	a run's number of points, and the deviation of channel c at point i
$g_{c,i}(u)$, $G_{c,i,k}$	a hypothesis's predicted shift, and the response to nuisance component k
μ_i, $\mathbf{1}_F$	a flow term, and one in a graph that carries flow terms
s_c, $\tilde{s}_c$, s_E	a channel's point spread, its row spread, and the balance's point spread
$\mathcal{C}_s$, $\mathcal{J}_s$, n_b, n_ν, n_p	a set's reading channels, joint-offset channels, other offsets, nuisance rows and prior rows
ζ_j, m_c, $\sigma_{b,c}$	a whitened joint offset, and an offset's prior mean and spread
t_k, c	the comparator's standardized residual on feature k and its threshold
y_k, $\hat{y}_k$, RMSE_k	a comparator feature, its reference-model prediction and that model's root-mean-square error
T_{rc}, T_{ci}, T_{co}, c_p, $\dot{m}_{cd}$	condensing, entering and leaving condenser water temperatures; water specific heat; condenser water mass flow
n_{10}, n_{01}	the discordant counts of a paired comparison (first method only right; second only right)
Δ	the difference in merged run-level accuracy, first method minus second

References

[Comstock and Braun 1999] Comstock, M. C., Braun, J. E. (1999). Development of Analysis Tools for the Evaluation of Fault Detection and Diagnostics for Chillers. ASHRAE Research Project 1043-RP Final Report, Report 4036-3, Ray W. Herrick Laboratories, Purdue University, West Lafayette, IN.

[Katipamula and Brambley 2005] Katipamula, S., Brambley, M. R. (2005). Methods for fault detection, diagnostics, and prognostics for building systems — a review, Part I. HVAC&R Research 11(1), 3–25. doi:10.1080/10789669.2005.10391123.

[Reddy 2007] Reddy, T. A. (2007). Application of a generic evaluation methodology to assess four different chiller FDD methods (RP-1275). HVAC&R Research 13(5), 711–729. doi:10.1080/10789669.2007.10390979.

[Zhou et al. 2009] Zhou, Q., Wang, S., Xiao, F. (2009). A novel strategy for the fault detection and diagnosis of centrifugal chiller systems. HVAC&R Research 15(1), 57–75. doi:10.1080/10789669.2009.10390825.

[Zhao et al. 2013] Zhao, Y., Wang, S., Xiao, F. (2013). A statistical fault detection and diagnosis method for centrifugal chillers based on exponentially-weighted moving average control charts and support vector regression. Applied Thermal Engineering 51(1–2), 560–572. doi:10.1016/j.applthermaleng.2012.09.030.

[Zhao, Xiao and Wang 2013] Zhao, Y., Xiao, F., Wang, S. (2013). An intelligent chiller fault detection and diagnosis methodology using Bayesian belief network. Energy and Buildings 57, 278–288. doi:10.1016/j.enbuild.2012.11.007.

[Han et al. 2011] Han, H., Gu, B., Hong, Y., Kang, J. (2011). Important sensors for chiller fault detection and diagnosis (FDD) from the perspective of feature selection and machine learning. International Journal of Refrigeration 34(2), 586–599. doi:10.1016/j.ijrefrig.2010.08.011.

[Clopper and Pearson 1934] Clopper, C. J., Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 26(4), 404–413. doi:10.1093/biomet/26.4.404.

[McNemar 1947] McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12(2), 153–157. doi:10.1007/BF02295996.

[Holm 1979] Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6(2), 65–70.

[Efron 1979] Efron, B. (1979). Bootstrap methods: another look at the jackknife. The Annals of Statistics 7(1), 1–26. doi:10.1214/aos/1176344552.

[Hotelling 1931] Hotelling, H. (1931). The generalization of Student's ratio. The Annals of Mathematical Statistics 2(3), 360–378. doi:10.1214/aoms/1177732979.

[Schölkopf et al. 2001] Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. Neural Computation 13(7), 1443–1471. doi:10.1162/089976601750264965.

[Nosek et al. 2018] Nosek, B. A., Ebersole, C. R., DeHaven, A. C., Mellor, D. T. (2018). The preregistration revolution. Proceedings of the National Academy of Sciences 115(11), 2600–2606. doi:10.1073/pnas.1708274114.

[Marquez 2026] Marquez, R. (2026). Identifiability, validation and acceptance testing of a physics-structured Bayesian fault-diagnosis model for a water-cooled centrifugal chiller, from normal operation under a pre-registered protocol. Technical report (the companion report).