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

Ricardo Marquez
rmarquez123@gmail.com

September 2026

Abstract

Model-based fault diagnosis of chillers rests on whether the model reproduces normal operation, which is seldom tested before faults are scored. This report asks what normal operation alone settles about a physics-structured Bayesian diagnosis model of a laboratory water-cooled centrifugal chiller, before any fault run is read. This is an inference study: the model was fitted to seventeen normal runs and tested by cross-validation over held-out days of testing, its acceptance test, identifiability predictions and fallback were fixed in a registered protocol beforehand, and the inference was validated on plants simulated from the model. The registered acceptance test of the refrigerant-side exchanger model **failed**: on the evaporator approach and on the subcooling the cross-validated error exceeded the limit of 1.25 times the error expected of a correct model, reaching 1.45 times it, while the condenser approach passed. Writing the compressor efficiency against the refrigerant lift rather than against the water temperatures cut the power error by a factor of 2.41. On the real normal runs, condenser fouling, non-condensable gas and a refrigerant leak cannot be told apart, and the predicted rate of calling an untroubled machine faulty rises from 0.074 with every sensor to about 0.4 without the refrigerant pressures. Self-recovery and calibration on simulated plants pass, so the shortfalls lie in what the model represents and what the measurements carry, not in the computation. The fault runs are therefore diagnosed in a companion report under the registered fallback, with the physics-based claims restricted to the fault classes the machine demonstrably expresses. The study covers one machine and reads no fault run; the computation on normal operation takes about 2.4 hours.

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 cooling season and that an operator has no direct instrument for. Automated fault detection and diagnosis — fault detection and isolation in the control literature, deciding that something is wrong and then which fault it is — sets out to read those conditions from the sensors a chiller already carries, and the difficulty is that the readings move for innocent reasons — a different load, a different 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 instrumented, run over a grid of loads and water temperatures, and then run again with seven faults introduced deliberately at graded severities, each against a matched normal run.

Methods proposed on that data set fall into three groups. Rule-based methods compare each reading against a regression of normal operation and match the pattern of signed residuals against a table of expected symptoms; Reddy (2007) evaluates four such methods on these runs under a common protocol and reports how sharply their performance falls with severity. Data-driven classifiers learn the fault label directly from the readings, which makes them strong on the data they were trained on and 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, including Zhou, Wang and Xiao (2009), and Zhao, Wang and Xiao (2013) report a statistical method on them. Nothing in this report is a first of its kind and nothing here claims to be. Two recurring evaluation hazards shape what follows. The first is that a model's adequacy on normal operation is usually asserted rather than tested, so a diagnosis failure cannot afterwards be attributed between the model and the data. The second is leakage: a benchmark scored by splitting measured points rather than whole runs lets a method recognize the run it is being asked about, and a method tuned while its fault-run scores are visible cannot be evaluated on those scores at all.

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 a single magnitude freed for each candidate fault so that competing explanations of one run can be weighed against one another. Section 3 states the framework and each of these constructs precisely.

This report is the first of two on that model, and it reads no fault run. It is an inference study: readings with a stated scatter and priors on every unknown are combined into a posterior, and competing explanations of a run are compared by their evidence. Its subject is what the model and the seventeen normal runs settle on their own, before any fault is scored: whether the refrigerant-side model reproduces normal operation well enough to carry a diagnosis, which candidate laws the normal runs select, which faults the instrumentation can tell apart at all, how often the model would call an untroubled machine faulty, and whether the inference recovers what it should on plants simulated from the model itself. Every threshold, draw count and pass rule below was fixed in a protocol file committed on 2026-09-22, before the computation ran, and the report states its results whichever way they came out: the acceptance test failed, and the fallback that failure triggers governs the companion report. Section 2 describes the machine and the measurements, Section 3 the model 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. Two water loops connect it to the laboratory: chilled water carries the cooling load from a test facility into the evaporator, and condenser water carries the rejected heat away. The machine, its two loops and every sensor this report uses are shown in Figure 1, with the shading that marks the sensor set at which each reading disappears.

Three temperature differences carry most of what the refrigerant side reveals. The *evaporator approach* is the leaving chilled-water temperature minus the saturated refrigerant temperature implied by the measured evaporating pressure; the *condenser approach* is the saturated refrigerant temperature implied by the measured condensing pressure minus the leaving condenser-water temperature; the *subcooling* is that same condensing temperature minus the measured temperature of the liquid leaving the subcooler. Each is a few kelvin, and each grows when the corresponding heat exchanger loses conductance. The oil-sump temperature is the fourth channel the model predicts. Converting a measured pressure to a saturation temperature uses a frozen property map for the refrigerant, which reproduces the benchmark's own recorded pressure and temperature pairs with a residual standard deviation of 0.0380144 K over the 970 of 972 recorded pairs it was fitted on, the remainder being trimmed as outliers.

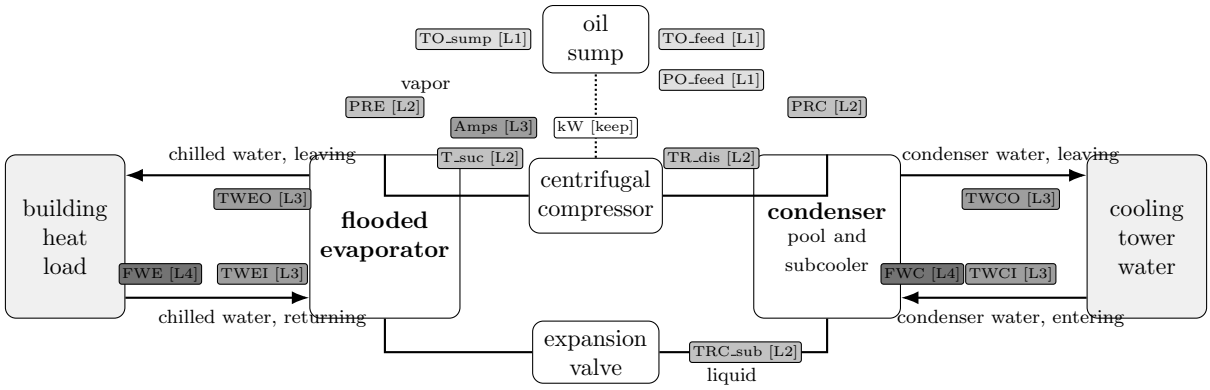
The committed extract holds 59 runs. Seventeen of them are normal runs usable by this report, and 31 are fault runs with a recorded severity; the fault runs are a population to be diagnosed in the companion report and contribute nothing here. Each run is a sequence of steady states on a grid of loads and water temperatures, 27 distinct operating conditions in all. The normal runs group into day blocks — maximal stretches of normal testing whose dates are at most one day apart — and there are 8 of them after the first, which predates a correction to the condenser-water flow measurement made in September 1999 and is excluded from every fit and every table in this report. The blocks and their 456 measured points are set out in Table 2, a count that already excludes one point removed by name: a run comment records a condenser-water flow reduction of about five percent at one test of one run, half the mildest deliberate flow fault, and that point is removed from every fit and every table, by comment and never by value. A screen for sensors reading the same quantity twice rejected 2 readings; no other screen is applied. Whenever a run must be compared with untroubled operation it is paired with the nearest earlier normal run, never across the flow-measurement correction.

A conservation audit of the same extract, reported separately (Marquez 2026), reproduced the recorded computed quantities of the normal runs at 486 measured points with zero disagreements, which is why the recorded channels are taken at face value here.

The same measurements are read at five nested sensor sets, from the laboratory instrumentation down to a list a building-automation system would hold; Table 1 defines them. Removing a raw sensor removes every recorded quantity computed from it, so the sets fall from 38 channels to 5. They are this report's way of asking which conclusions survive on a real building's instrumentation.

Assumptions. The following are held fixed throughout.

- Temperatures are carried in kelvin and powers in watts; each table states the unit of each column, and severities keep the benchmark's own units.
- The refrigerant property map is frozen and is never refitted per run.



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.

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 evaporator approach is the leaving chilled-water temperature minus the saturated refrigerant temperature implied by the evaporating pressure; the condenser approach is the analogous difference on the condenser side; the subcooling is the condensing temperature minus the temperature of the liquid leaving the subcooler. Those three differences, and the oil-sump temperature, are the quantities whose model is tested here.

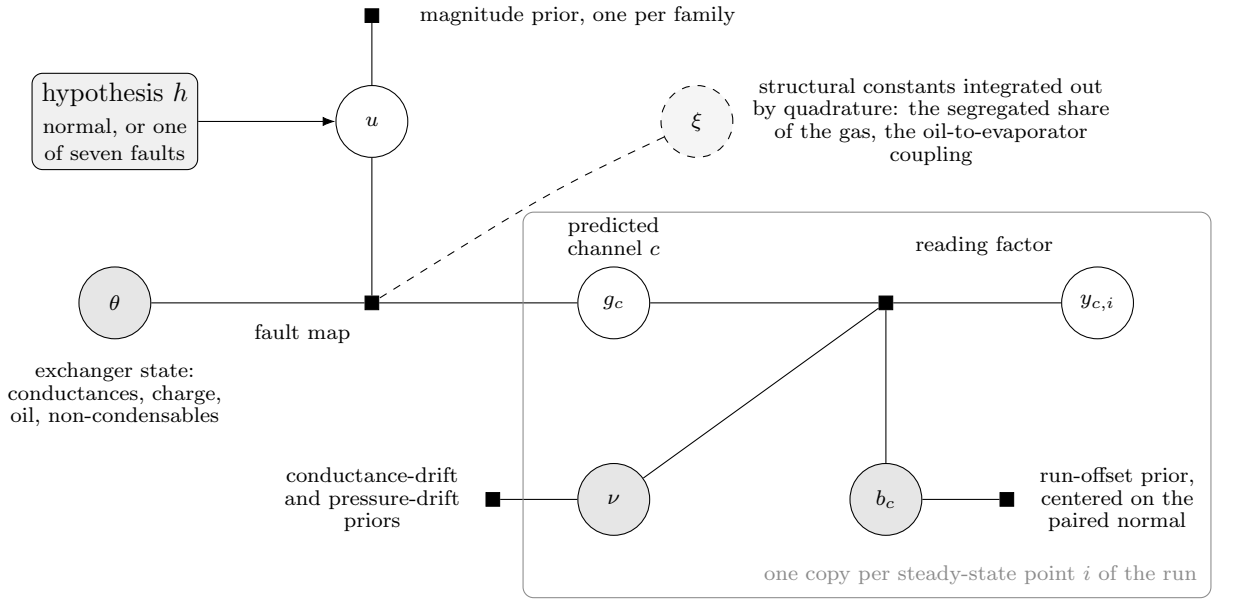
- A run's steady states are repeated measurements of one machine state, with the scatter split into a day-block component, a run component and a point component.
- Every structural constant that normal operation does not identify is carried over a registered range rather than taken as exact: two are integrated out of the evidence by quadrature, and the others widen the magnitude prior of the family whose severity map they enter; all of them appear among the constants of Table 22.
- Normal runs are normal by the benchmark's own design: no fault was introduced in them.

The model these assumptions describe is set out as a factor graph in Figure 2: the unknowns as circles, the relations that tie them as black squares, and, dashed, what is integrated out.

3 Methods

3.1 Overview

The model is built in three layers: a description of the machine in normal operation, fitted to the normal runs; a probabilistic layer that turns that description, the sensor noise and the priors into a posterior; and a hypothesis layer that displaces the description by one fault at a time and compares the explanations. The order of the computation is laid out in Figure 3, one box per step with the subsection that states it, and a dashed line marks where the protocol was fixed. The normal runs are grouped into day blocks and held out one block at a time; on each fold the refrigerant-side laws (Section 3.2) and the compressor-power law (Section 3.3) are refitted, in the forms the normal runs select (Section 3.4), with the scatter split as in Section 3.5. The hypotheses and their priors



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.

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 TEA, TCA, TRC_sub, TO_sump, Tsh_suc, compressor_power, energy_balance, FWE, FWC. 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. Normal operation fixes everything outside the hypothesis layer; the hypothesis layer is what this report tests.

Table 1: The five sensor sets. Each removes a group of instruments and every recorded quantity computed from them, so the sets are nested. Beyond these, 11 recorded columns are given to no method at any set (test indices, set points, valve positions and the fault flag itself), and 17 channels of the test facility's own conditioning loops are dropped throughout, because they describe the laboratory and not the machine.

Set	What it stands for	Channels [count]	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	..., `TRC_sub`, `T_suc`, `TR_dis`, `TWEI`, `TWEO`, `TWCI`, `TWCO`, `Amps`
L4	the same without flow meters	5	..., `TR_dis`, `TWEI`, `TWEO`, `TWCI`, `TWCO`, `Amps`, `FWE`, `FWC`

Table 2: The normal runs, grouped into day blocks. A day block is a maximal stretch of normal testing whose dates are at most one day apart. Block 0 predates a correction to the condenser-water flow measurement and is excluded from every fit and every table; the remaining 17 runs, 456 measured points in all after the one recorded point exclusion, are the population of this report. Every cross-validation below holds out one block at a time.

Day block [index]	Runs	Measured points [count]	Role
0	normal dpv	—	excluded everywhere
1	near normal3, near normal1, near normal2, normal	108	held out once
2	normal1	27	held out once
3	normal2	27	held out once
4	normal r	27	held out once
5	normal r1, normal eo, normal b	81	held out once
6	normal cf	27	held out once
7	normal cf2, normal cf3, normal cf4, normal cf5, normal cf6	132	held out once
8	normal nc	27	held out once

(Section 3.6) are weighed in the probabilistic model of Section 3.7, whose rows Section 3.8 writes out in full. The held-out errors feed the acceptance test (Section 3.9); the per-fold models give the fault signatures for identifiability and detectability (Section 3.10) and the simulated normal runs for the predicted false alarms (Section 3.11); a separate branch validates the inference on simulated plants (Section 3.12). The protocol and the metrics close the section.

3.2 The refrigerant-side model

Each heat exchanger is described by a conductance — the rate at which it moves heat per kelvin of temperature difference — which is not constant but varies with load and with the water temperatures. The evaporator's conductance is written on a log scale as a linear combination of a small set of features of the operating point,

$$\ln UA_e = \sum_j a_j \, \phi_j\big(Q,\ T_{ewo},\ T_{cwi}\big), \tag{1}$$

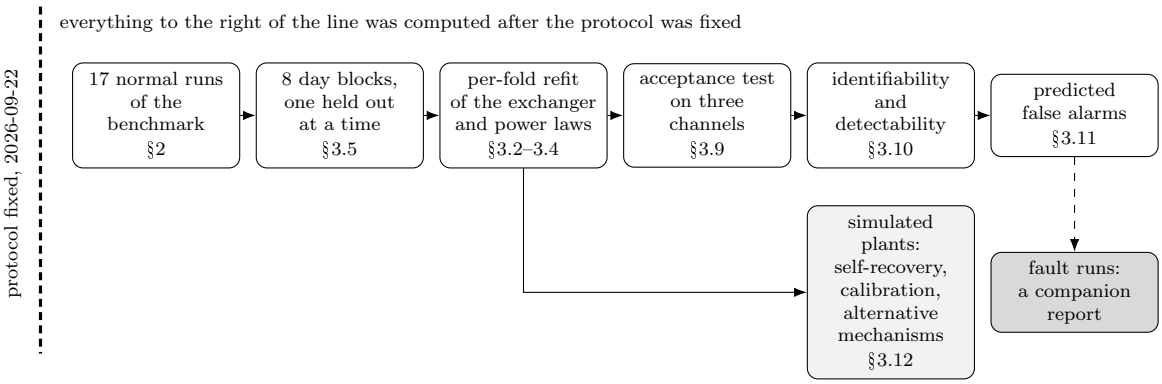


Figure 3: The order of the computation, one box per step with the subsection that states it. Normal runs are grouped into day blocks; each block is held out in turn and the exchanger and power laws are refitted on the rest, which gives a cross-validated error per channel. That error is compared with the error a correct model would show given the measured scatter — the acceptance test. The same per-fold models give the fault signatures from which identifiability and detectability follow, and the rate at which the model calls a simulated normal run faulty. A separate branch validates the inference on simulated plants. The dashed vertical line marks the date the protocol was fixed: every quantity in this report was computed after it. The fault runs are diagnosed in a companion report and hold no content here.

with Q the cooling load, T_{ewo} the leaving chilled-water temperature and T_{cwi} the entering condenser-water temperature. The condenser and the subcooler carry laws of the same shape. The feature sets were chosen by cross-validation on these same normal runs, before the protocol was fixed, and the protocol records the choice and its consequence: a test of the model on the runs that selected it is generous, so the acceptance test of Section 3.9 is also run in a stricter form that repeats the selection inside every fold. The features the evaporator uses are `1 + ln(Q / Q_ref) + T_ewo - 45 F + T_cwi - 75 F + ln(Q / Q_ref) (T_ewo - 45 F)`. The conductances of (1) predict the three approach temperatures; a two-regime charge inventory relates the refrigerant mass to the pool level and the flooded area; an oil law ties the sump temperature to the oil inventory; and a dry-zone law gives the suction superheat.

3.3 The compressor-power law

The compressor's power is written as the reversible work of lifting the evaporator load from the evaporating temperature to the condenser shell's saturation temperature, divided by an efficiency that varies with the load and with that lift, plus a constant power,

$$P = \frac{Q\,(T_s - T_e)}{T_e\,\eta} + P_0, \tag{2}$$

$$\eta = \eta_0\big(1 + e_1 x + e_2 x^2 + e_3(T_s - T_e - T_\lambda)\big), \tag{3}$$

$$x = \ln(Q/Q_{\mathrm{ref}}), \tag{4}$$

where Q is the evaporator load, T_e the evaporating temperature and T_s the condenser shell's saturation temperature, both predicted by the fold's exchanger model from the water side, $Q_{\mathrm{ref}} = 316.5$ kW the reference load and $T_\lambda = 25$ K the reference lift. The five coefficients of (2) and (3) — the efficiency η_0 at the reference load and lift, its relative slopes e_1, e_2 in the log load x of (4) and e_3 in the lift, and the constant power P_0 — are fitted per fold by Gauss–Newton in the power itself on the training normals. At the sensor sets without refrigerant pressures the exchanger model

Table 3: Candidate laws compared on the normal runs by holding out one day block at a time, 8 folds over the same 456 measured points the acceptance test of Section 3.9 runs on. The error is the root-mean-square difference between measurement and prediction on the held-out block, pooled over blocks, in the unit shown. The selected law is the one with the smallest such error in its group, and it is the law carried into the model; the rebuilt lift is not a competitor but the fallback used where the refrigerant pressures are unavailable, and it is shown here beside the forms it replaces.

Quantity	Candidate law	Unit	Cross-validated error [in the unit shown]	Outcome
compressor power	efficiency against the refrigerant lift	kW	0.944495	selected
compressor power	efficiency against the water temperatures	kW	2.2775	—
compressor power	lift rebuilt from design approaches	kW	2.28919	—
oil-sump temperature	refrigerant state and water conditions	K	0.584653	selected
oil-sump temperature	water conditions only	K	0.790065	—
oil-feed temperature	the predicted sump temperature	K	0.722814	selected
oil-feed temperature	water conditions only	K	0.830912	—
suction superheat	dry zone on the measured approach	K	0.139795	selected
suction superheat	dry zone on the predicted approach	K	0.22944	—
suction superheat	linear in the water conditions	K	0.220345	—

cannot be fitted, and the two saturation temperatures are reconstructed from design approaches added to the measured water temperatures; L2, L3, L4 use that form.

3.4 Selection of the laws on normal operation

Each candidate law was compared with its alternatives by holding out one day block at a time, fitting on the rest and measuring the error on the held-out block. Table 3 gives the comparisons: the candidate forms of the compressor-power law, of the oil-sump and oil-feed laws and of the suction superheat, each with its cross-validated error. The same normal runs served both to select the laws and, later, to test the selected model; the protocol states this plainly, and Section 3.9 reports the stricter check in which the selection is repeated inside each fold.

3.5 The scatter of normal operation

Repeated measurements of the same machine differ for three reasons, and the model separates them: a day-block component, which moves a whole day's readings together as the machine and the laboratory settle differently; a run component within a block; and a point component from the instruments themselves. Their variances are estimated from the normal runs by the classical decomposition of a nested design (Henderson 1953), so that a reading's total variance is

$$\mathrm{Var}(y_{c,i}) = \sigma^2_{B,c} + \sigma^2_{R,c} + \sigma^2_{P,c}. \tag{5}$$

Table 4 gives the three components of (5) per channel. Two further terms are carried per run rather than per point: a conductance drift on each loop, which lets a run's exchangers sit slightly away from the fitted law without that being read as a fault, and a discrepancy attached to the operating condition, which absorbs the part of the model error that repeats at the same load and water temperatures. The values of the per-condition discrepancy are not published by design: publishing them would amount to publishing a second, unfitted model of the same runs.

Table 4: The scatter of normal operation, split three ways and estimated without any model of a fault, from 17 runs in 8 day blocks. Standard deviations, in kelvin. The day-block component is what a model fitted on other days cannot know, and it is what the acceptance test's expectation in Section 3.9 is built on.

Channel	Day block [K]	Run [K]	Point [K]	Total [K]
TCA — condenser approach	0.120115	0.0779073	0.168412	0.221043
TEA — evaporator approach	0.148312	0.0949854	0.141936	0.226196
TO_sump — oil-sump temperature	0.289123	0.261437	0.158499	0.420789
TRC_sub — subcooling	0.0959519	0.0207247	0.166038	0.192886

3.6 Hypotheses and priors

There are eight hypotheses: normal, and one for each of the seven fault families of the benchmark. Each fault frees exactly one magnitude M, constant within a run, carried on a log scale as $u = \ln M$ with a prior

$$u \sim \mathcal{N}\big(m_h,\ s_h^2\big). \tag{6}$$

The prior of (6) is set from the severity range the benchmark quantifies for that family, widened by the uncertainty in the map from severity to magnitude; Table 7 sets out both parts. Three constants of that map are not identified by normal operation at all, because normal runs contain none of the fault in question: how much of the condenser's thermal resistance sits in the water film, taken as 0.65 over a range the protocol fixes; how much of a nitrogen charge sits as a segregated pocket blanketing the tube bundle rather than mixed into the vapor, taken as uniform over its whole range; and how strongly excess oil in the sump couples into the evaporator, taken as a point mass at no coupling plus a uniform spread above it. The last two, the segregated share and the oil coupling, are integrated out of the evidence of the non-condensable and the excess-oil hypotheses rather than fixed,

$$Z_h = \int Z_h(\xi)\, p(\xi)\, d\xi, \tag{7}$$

by quadrature over the constant's own prior. The water-film share enters differently. When a fraction f_p of the condenser tubes is plugged, the open area falls to $1 - f_p$ while the open tubes carry the whole flow faster, so the film's share ψ of the thermal resistance falls with the water velocity and the magnitude of condenser fouling is

$$M = -\ln \frac{1 - f_p}{\psi\,(1 - f_p)^{0.8} + 1 - \psi}, \tag{8}$$

with 0.8 the exponent of the film's conductance in the water velocity. The share is held at its central value in (8), and its registered range $[\psi_-, \psi_+] = [0.4, 0.9]$ widens the condenser-fouling prior instead: the magnitudes it gives at the mildest quantified severity f_1 set a factor whose logarithm is added in quadrature to the spread of the severity range,

$$\Gamma_\psi = \sqrt{M(f_1; \psi_-)/M(f_1; \psi_+)}, \tag{9}$$

$$s_{\mathrm{cf}}^2 = s_{\mathrm{grid}}^2 + (\ln \Gamma_\psi)^2, \tag{10}$$

where s_{grid} is the half-width, on the log scale, of the magnitudes at the two ends of the quantified severity range; the factor of (9) and its logarithm are the structural columns of the condenser-fouling

row of Table 7. The composition of the non-condensable prior in (7) is built over 456 operating conditions. Hypotheses whose posterior probabilities agree to a relative tolerance of 1e-06 are one answer to the data, and such ties are ordered by name, never by round-off.

3.7 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,

$$p(z \mid y) \propto \prod_k \exp\!\big(-\tfrac{1}{2}\, w_k\, r_k(z)^2\big). \tag{11}$$

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:

$$C(z) = \tfrac{1}{2} \sum_k w_k\, r_k(z)^2, \tag{12}$$

$$z^* = \arg\min_z C(z), \tag{13}$$

$$z \leftarrow z - (J^\top W J)^{-1} J^\top W\, r(z) \quad \text{until } \|J^\top W\, r(z)\| < \epsilon. \tag{14}$$

Near the minimum the cost (12) 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:

$$p(z \mid y) \approx \mathcal{N}\big(z^*, \Lambda^{-1}\big), \tag{15}$$

$$\Lambda = J^\top W J, \tag{16}$$

$$\sigma_i = \sqrt{(\Lambda^{-1})_{ii}}. \tag{17}$$

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,

$$\log Z = -C(z^*) + \tfrac{d}{2} \log 2\pi - \tfrac{1}{2} \log\det \Lambda + \sum_k \tfrac{1}{2} \log \frac{w_k}{2\pi}. \tag{18}$$

Differences of $\log Z$ between hypotheses, from (18), 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.

3.8 The residual rows of one hypothesis

The graph of Section 3.7 is assembled once for each run, sensor set and hypothesis, and its rows are written out here in full. A run has I steady-state points, indexed by i; a simulated run of this report carries one point per operating condition, 27 in all. 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,

$$\tilde{y}_{c,i} = y_{c,i} - \hat{y}_{c,i} - \delta_{c,t_i}, \tag{19}$$

$$\tilde{y}_{E,i} = \frac{Q_{C,i} - P_i}{D_i} - 1, \tag{20}$$

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 (20),

$$r_{E,i} = b_E - \tilde{y}_{E,i} + \mu_i, \tag{21}$$

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,

$$r_{F,i} = \mu_i - (1 + b_E)\big(s_E(u) - 1\big) + q_i\big(s_C(u) - 1\big), \tag{22}$$

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 (22) 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,

$$r_{c,i} = \tilde{y}_{c,i} - b_c - \sum_k G_{c,i,k}\,\nu_k - g_{c,i}(u), \tag{23}$$

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,

$$\tilde{s}_c^2 = s_c^2 + \frac{1}{I}\sum_i \frac{s_c^2}{n_{t_i}}, \tag{24}$$

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,

$$r_{b,c} = b_c - m_c, \tag{25}$$
$$r_{\zeta,j} = \zeta_j, \tag{26}$$
$$r_{\nu,k} = \nu_k, \tag{27}$$
$$r_u = u - m_h, \tag{28}$$

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,

$$b_c = m_c + \sum_j L_{cj}\,\zeta_j. \tag{29}$$

A nuisance component is carried at unit spread, its scale being in $G_{c,i,k}$, and the magnitude row is the prior of (6). The offset prior is centered on the paired normal's own offset m_c, and for a fold with B training day blocks its variance is

$$\sigma_{b,c}^2 = \frac{B-1}{B-3}\Big(2\big(s_{B,c}^2 + s_{R,c}^2\big) + \frac{s_c^2}{I_0}\Big) - \omega_c, \tag{30}$$

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 5 lists every row family with its residual, unit, spread, weight and multiplicity, and Table 6 gives the multiplicities at each sensor set. For hypothesis h at set s the residual vector r, the Jacobian J and the weight matrix W of (14) have the dimensions

$$\dim r = I\,\big(1 + |\mathcal{C}_s| + \mathbf{1}_F\big) + n_p, \tag{31}$$
$$n_p = n_b + |\mathcal{J}_s| + n_\nu + \mathbf{1}[h \neq \text{normal}], \tag{32}$$
$$d = n_p + I\,\mathbf{1}_F, \tag{33}$$

so that J is $\dim r$ by d and W is the diagonal $\dim r$ by $\dim r$ matrix of the weights; n_p counts the prior rows, $\mathbf{1}[\cdot]$ is one when the statement in brackets holds, and the d unknowns are one per prior row and one flow term per point of a flow-term graph.

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 (1), the charge inventory, the oil law and the power law (2) 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 (24), 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.

Several calculations of this report lie outside the graph and contribute no row to it: the per-fold fits of the conductance, superheat, oil and power laws, each by unweighted least squares on the training normals, law by law; the scatter components and the condition means δ, estimated by moments on the same normals; the cross-validation and acceptance test of Section 3.9; the signature angles of Section 3.10; and the drawing of simulated runs. The rows are used in two ways. The detectable magnitude of Section 3.10 reads the information $I(M)$ from the maximum-a-posteriori solve and its Laplace curvature, (14) and (16), on the expected run at the mildest severity. The predicted false alarms and the validation programs evaluate each hypothesis's evidence on the same 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; the validation programs assemble the same families on a synthetic plant with the scatter each program states and no per-condition discrepancy.

3.9 The acceptance test

A diagnosis rests on the claim that the model reproduces a machine with nothing wrong with it. The protocol fixed a test of that claim, on the three refrigerant-side channels the diagnosis leans on, before any of it was computed. Each day block k is held out in turn, the model is refitted on the rest, and its prediction $\hat{y}_{c,i}^{(-k)}$ of channel c at every held-out point i is compared with the measurement $y_{c,i}$. The squared differences are pooled over all N held-out points of all blocks into the cross-validated error,

$$\mathrm{CV}_c^2 = \frac{1}{N} \sum_{k} \sum_{i \in k} \left(y_{c,i} - \hat{y}_{c,i}^{(-k)} \right)^2, \tag{34}$$

in kelvin, where $\sum_{i \in k}$ runs over the points of block k; the same sum restricted to one block is that block's error. The error of (34) is compared with the error a *correct* model would show, which is not zero: the held-out block carries its own day-level offset, which the fit could not have seen, and the fitted mean carries the error of the blocks it was fitted on. For a fold with B training day blocks and n training runs, a correct model's expected squared error at a held-out point is

$$E_c^2 = \sigma_{B,c}^2 \Big(1 + \tfrac{1}{B} \Big) + \sigma_{W,c}^2 \Big(1 + \tfrac{1}{n} \Big), \tag{35}$$

with $\sigma_{B,c}$ the between-block and $\sigma_{W,c}$ the within-block standard deviation of the channel, both in kelvin and both estimated from the normal runs split into day blocks; (35) is averaged over the held-out points in the same way as (34). The first term is the held-out block's own offset plus the error of a mean estimated from B blocks, and the second the same for the within-block scatter. The acceptance ratio is the quotient of the two,

$$R_c = \mathrm{CV}_c / E_c, \tag{36}$$

which is near one for a correct model with correctly stated scatter, and the pass rule bounds it by a fixed factor κ on every gated channel c:

$$R_c \le \kappa, \tag{37}$$

with $\kappa = 1.25$. The part of the error the scatter does not explain is reported alongside (37) as a bound in kelvin,

$$b_c = \sqrt{\max\left(0,\ \mathrm{CV}_c^2 - E_c^2\right)}, \tag{38}$$

Table 5: 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 (21), (22), (23) and (25). 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 (24); 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 6; 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, (30)
H_j	ζ_j	–	1	1	$\lvert\mathcal{J}_s\rvert$	joint offsets, (29)
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	Table 7

Table 6: The multiplicities of Table 5 at each sensor set, read from the fold trained on every usable normal. $|\mathcal{C}_s|$ counts the reading channels, $|\mathcal{J}_s|$ the channels whose offsets share the joint prior of (29), 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; *Rows at* $I = 27$ is the physics part of (31) for a simulated run of this report without a flow term. *Offset priors* is $n_b + |\mathcal{J}_s|$; (32) 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 [–]	Rows at $I = 27$ [count]	Offset priors [count]	s_E [–]	s_c, power [kW]
L0	8	6	3	≤ 1	9	—	243	9	0.0320569	0.567034
L1	7	5	3	≤ 3	8	—	216	8	0.0320569	0.567034
L2	3	0	4	≤ 2	4	—	108	4	0.0320569	1.07457
L3	3	0	4	≤ 2	4	—	108	4	0.0321964	1.07273
L4	1	0	2	≤ 2	2	{fwe, fwc}	54	2	0.0332907	1.0834

against the value (38) would take exactly at the limit. The gated channels are TEA, TCA, TRC_sub; the oil-sump channel is reported and not gated. The test is run in two forms: as fixed, with the conductance-law feature sets held at their chosen values, and in a stricter form that repeats the selection inside every fold. A point-level reference, which ignores the day-level component, is reported for comparison and governs nothing.

Knowledge available in advance. Two facts make the verdict partly foreseeable, and the protocol states both. The point-level ratios on all three channels had been measured on a larger normal population before the protocol was fixed (1.43, 1.23 and 1.42 are this report's recomputation of the same reference on its own population), and the day-block reference and the stricter multiplier were both chosen after those were seen and before the day-block statistic existed. Separately, the numerators of two of the three channels — the cross-validated errors themselves, though not the expectations they are divided by — had been computed while the model was being built. On those two channels the test therefore asks only whether the day-level component of the scatter is large enough to admit an error of a size already known. The third channel's numerator, and every numerator of the stricter form, had never been computed. Everything that had been computed on normal operation beforehand is enumerated in Table 20.

3.10 Identifiability and detectability

Two faults are confusable when they move the measured channels in nearly the same direction. For each family the model is displaced by that family's smallest quantified severity, and the resulting change in the four refrigerant-side channels is divided through by the measured scatter, which gives a signature s_f whose length is the number of standard deviations of reading the fault would produce,

$$s_f = \Sigma^{-1/2}\big(g(\theta_f) - g(\theta_0)\big), \tag{39}$$

$$\cos\vartheta_{fg} = \frac{s_f^\top s_g}{\|s_f\|\,\|s_g\|}. \tag{40}$$

The angle of (40) is compared with the angle the pair's own lengths would require before the two could be told apart; families that no comparison separates are merged into one class. A family

whose signature is shorter than the run-to-run spread of normal operation, 1 standard deviation, is marked as lying inside that spread. The protocol registered the classes it expected as values; a recomputation that disagrees is reported as a disagreement and changes no score.

Detectability asks the same question against noise rather than against another fault. Given the information $I(M)$ the readings carry about a family's magnitude, the smallest magnitude a test of size α and power $1-\beta$ resolves — the minimal detectable magnitude, the counterpart for a fault of the minimal detectable bias of state estimation — is

$$M_{\min} = \frac{z_{1-\alpha} + z_{1-\beta}}{\sqrt{I(M)}}, \tag{41}$$

evaluated at $\alpha = 0.05$ and power 0.8, and the magnitude (41) gives is converted back to the family's own severity unit for reporting.

3.11 Predicted false alarms

The model's own tendency to see a fault where there is none is predicted before any measurement is scored. For each held-out day block, normal runs are drawn from that block's own fitted description — readings with the measured scatter and no fault — and the model is asked for its most probable hypothesis: 200 draws per block and sensor set, 1,600 per set in all, from a fixed seed. Of the N_s draws at a sensor set s, the fraction answered with a hypothesis other than normal is the predicted false-alarm rate (the protocol's false-call rate),

$$\hat{f}_s = \#\{\, j : \hat{h}_j \neq \text{normal} \,\} \,/\, N_s, \tag{42}$$

where $\hat{h}_j$ is the most probable hypothesis for draw j and $\#\{\cdot\}$ counts the draws that satisfy the condition in the braces. The rate of (42) is evaluated per held-out block as well as pooled, and each block's rate carries an exact binomial interval (Clopper and Pearson 1934), an interval on a proportion computed from the binomial distribution itself rather than from a Gaussian approximation to it.

3.12 Validation on simulated plants

A diagnosis that misses a fault misses it for one of three reasons, and they call for different responses: the inference is implemented wrongly; the model does not represent the mechanism at work; or the measurements cannot settle the question at any implementation. Three programs separate them, all on plants simulated from the model and never fitted to the benchmark, at a single fixed seed. *Self-recovery* simulates each family at each quantified severity and asks whether the model recovers the family and whether its magnitude interval contains the truth; a failure here is an implementation failure. *Calibration* draws a magnitude from the prior of (6), simulates readings from it, and checks that the model's stated confidence matches how often it is right and that its intervals cover at their nominal rate, which is the simulation-based calibration check of Talts et al. (2018). The *alternative-mechanism* program simulates plants that do not follow the model's own mechanism — a different condenser pool geometry, a different valve behavior, a different distribution of non-condensable gas — against a matched control, and separates a representation gap from a limit of the measurements. Each program is run at the scatter measured on the real normal runs and again on a quieter plant.

3.13 Protocol and registration

Every number above — the multiplier 1.25, the gated channels, the expected classes, the draw counts, the seeds, the significance level and the power, the priors and the structural ranges —

was fixed in a file committed on 2026-09-22, before any of the computation of Section 4 ran and before any fault run had been read by any component of the method (Nosek et al. 2018). The same file fixes what happens if the acceptance test fails: the model's physics-based claims are then restricted to the classes the machine demonstrably expresses, and the confirmatory comparison of the companion report runs on a restricted population of 9 water-flow-fault runs together with the scored normal runs. One correction to the software was made after the protocol was fixed and before any result existed: the first complete computation stopped while writing its output, on a quantity that is legitimately undefined — the scatter of the oil channel at the sensor set that removes the oil sensor — and wrote nothing at all. The correction changed how undefined quantities are written out, and changed no registered value, no model and no measurement; it is itself dated and additive, and Table 19 carries the record. From that correction until the results existed, nothing was changed in the model, the protocol or the software.

3.14 Metrics

Every measure this report quotes is a numbered relation, and each is read against a stated reference. Grouped by the question of Section 1 each answers, they are as follows. *Adequacy of the model*: the acceptance ratio R_c of (36), dimensionless, near one for a correct model and read against the limit $\kappa = 1.25$ of (37); the misfit bound b_c of (38), in kelvin, read against the value it takes exactly at the limit; and the cross-validated error of (34), in kelvin for temperatures and in kilowatts for power, by which the candidate laws are compared, a law being preferred when it is lower in every held-out block and not merely on average. *Separability of the faults*: the signature angle of (40), in radians, read against the angle the pair's own magnitudes require, and the smallest detectable severity of (41), in the family's own unit, read against the mildest severity the benchmark quantifies. *Untroubled operation*: the false-alarm rate of (42), a fraction of draws with its exact binomial interval. *Soundness of the inference*, on simulated plants whose truth is known: for D draws, draw j simulated under hypothesis h_j with magnitude M_j and answered with the most probable hypothesis $\hat{h}_j$, its stated probability p_j and the stated magnitude interval $[L_j, U_j]$, the identification rate of family f over the set $\mathcal{D}_f$ of its own D_f draws, the coverage, the mean confidence and the accuracy are

$$\rho_f = \#\{\, j \in \mathcal{D}_f : \hat{h}_j \text{ is } f \,\} \,/\, D_f, \tag{43}$$

$$\gamma = \#\{\, j : L_j \le M_j \le U_j \,\} \,/\, D, \tag{44}$$

$$\bar{p} = \frac{1}{D} \sum_j p_j, \tag{45}$$

$$a = \#\{\, j : \hat{h}_j \text{ is } h_j \,\} \,/\, D. \tag{46}$$

All four are dimensionless fractions. The identification rate (43) is read against one, the coverage (44) against the nominal level of the stated interval, and the mean confidence (45) against the accuracy (46): a calibrated model states, on average, exactly the probability with which it is right, and the accuracy carries an exact binomial interval for that comparison.

4 Results

4.1 Adequacy of the refrigerant-side model

The acceptance test is summarized in Table 8, and Figure 4 breaks it down block by block. The table has one row per channel: the pooled cross-validated error of (34), the error expected of a

Table 7: Magnitude priors, one per fault family. Severities are percentages of the quantity named in the third column, and *Range* is the range the benchmark quantifies. Each fault hypothesis frees a single magnitude, carried on a natural-log scale, whose prior mean and standard deviation place the severity range the benchmark quantifies inside the prior's bulk. The last two columns describe the extra spread contributed by the structural constants that normal operation does not identify: how far the map from severity to magnitude may move, and the standard deviation of that movement on the log scale. A value printed as a bound is at the round-off floor.

Family	Fault	Severity, percent of	Range [%]	Prior mean [–]	Prior s.d. [–]	Struct. factor [–]	Struct. s.d. [–]
cf	condenser fouling	tubes plugged	6–45	-2.33275	1.25555	1.55879	0.443908
fwc	reduced condenser water flow	flow reduction	10–40	-1.46105	0.78932	—	—
fwe	reduced evaporator water flow	flow reduction	10–40	-1.46105	0.78932	—	—
rl	refrigerant leak	charge removed	10–40	-1.60944	0.859996	1.66372	0.509058
ro	refrigerant overcharge	charge added	10–40	-1.60944	0.780323	1.43104	0.358401
eo	excess oil	excess oil	14–73	1.03434	1.07807	2	0.693147
nc	non-condensable gas	condenser volume	1–5.6	1.17422	1.04864	1.81856	0.598044

correct model from (35), the ratio (36) of the two under the test as fixed, under the stricter form and under the point-level reference, the misfit bound of (38), and the value that bound takes at the limit. The figure has one panel per channel; each dot is a held-out day block, sized by how many points it contains, and the three horizontal lines are the pooled error, the expectation and the limit.

The test failed. The evaporator approach came in at 1.26425 times its expectation and the subcooling at 1.4481, against the limit of 1.25 in (37); the condenser approach passed at 1.1302. The stricter form, which repeats the feature selection inside every fold, gives 1.26425, 1.1324 and 1.41455: the same two channels fail. Read in kelvin rather than as a ratio, the evaporator approach carries 0.178767 K of unexplained error where 0.173331 K is admitted, and the subcooling 0.188644 K where 0.135084 K is admitted; on the condenser approach the unexplained part, 0.109763 K, is comfortably inside its allowance of 0.156315 K. The oil-sump channel, reported and not tested, stands at 1.56636, or 1.56541 under the stricter form.

Figure 4 shows where the failure sits; this is a description of the same numbers and not a second test. On the evaporator approach two blocks carry most of it, block 1 at 0.413185 K and block 6 at 0.411221 K, against a pooled 0.292179 K. On the subcooling the largest single block is block 4 at 0.371956 K. The pattern is day-to-day rather than uniform: the refrigerant-side conductances sit differently from one day of testing to the next by more than the laws of (1) allow them to.

The verdict was partly known in advance, as Section 3.9 sets out: the numerators of the evaporator and condenser approaches had been computed before the protocol was fixed, so on those two channels the test asks only whether the day-level component of the scatter is large enough to admit an error already known. The subcooling's numerator, and all three numerators of the stricter form, had not been computed; the subcooling fails on both counts.

The three forms of the ratio are set side by side because each answers a different objection to the verdict. The point-level reference ignores the day-to-day component of the scatter and so expects a smaller error; the test as fixed adds that component through (35); the stricter form also removes the advantage of having chosen the feature sets on the runs that test them. That the same two channels fail under the second and the third means the verdict rests neither on a reference chosen to be harsh nor on features chosen to flatter. The comparison demonstrates that a cross-validated error read against a stated scatter, with the limit fixed in advance, localizes a model's inadequacy to particular channels and particular days, which a single pooled goodness-of-fit number would not.

Table 8: The acceptance test. *Error* is the pooled cross-validated root-mean-square difference between measurement and prediction; *Expected* is the error a correct model would show given the measured scatter of that channel. The three ratio columns are error over expectation under the test the protocol fixed, under the same test with the model's feature choice repeated inside every fold, and under a point-level reference reported for comparison. A channel passes when its ratio is at most 1.25. *Misfit* is the part of the error that the measured scatter does not explain, and *Allowed* is how large that part may be at the limit. The channels are the evaporator approach, the condenser approach, the subcooling and the oil-sump temperature; the last is reported and not tested.

Channel	Error [K]	Expected [K]	Ratio, as fixed [–]	Ratio, stricter [–]	Ratio, point level [–]	Misfit [K]	Allowed [K]	Verdict
TEA	0.292179	0.231108	1.26425	1.26425	1.42931	0.178767	0.173331	fail
TCA	0.235556	0.20842	1.1302	1.1324	1.22771	0.109763	0.156315	pass
TRC_sub	0.26082	0.180112	1.4481	1.41455	1.42303	0.188644	0.135084	fail
TO_sump	—	—	1.56636	1.56541	—	—	—	reported

4.2 The compressor-power law

Figure 5 plots the cross-validated error of the two candidate power laws, block by block: circles for the lift law of (2) and (3), whose efficiency is written against the refrigerant lift, squares for the law of Gordon and Ng (1995) written in the water temperatures, with each form's pooled error as a dotted line. Writing the efficiency against the lift is better in eight of eight blocks, and pooled it gives 0.944495 kW against 2.2775 kW, a reduction by a factor of 2.41. The improvement is not an artifact of fitting: the lift law carries five coefficients against the water-temperature law's three, but both are scored only on blocks they were not fitted on, so the extra coefficients earn their place only by what they predict on days not seen. Where the refrigerant pressures are unavailable and the lift must be reconstructed from design approaches added to the water temperatures, the same cross-validation gives 2.28919 kW, which is 2.4 times the measured-lift form and close to the water-temperature form: most of what the lift contributes is lost when it has to be inferred. The computations that could not be carried out at those sensor sets are counted in Table 9, all for the same reason — with no condensing temperature there is no exchanger model to differentiate — and no result of this report depends on them.

The two forms are compared because both are in use for chiller power and they differ in the temperatures the efficiency is written in: the saturation temperatures that set the lift in (3), or the water temperatures. The gap between them is what writing the law on the thermodynamic driver of the work, the refrigerant lift, buys over writing it on the water temperatures that set the lift only indirectly; that the gap closes when the lift has to be rebuilt from design approaches shows the gain comes from saturation temperatures predicted by an exchanger model fitted to the measured refrigerant pressures, and not from the form of the law alone. The comparison demonstrates that cross-validation over held-out day blocks selects a law by what it predicts on days it has not seen, not by how well it fits the days it was given.

4.3 Identifiability on normal operation

The angles of (40) between the fault signatures of (39) at the laboratory instrumentation are arranged as a matrix in Figure 6 and tabulated pair by pair in Table 10. Each cell of the matrix holds the angle between two families' signatures and, below it, the angle that pair would need;

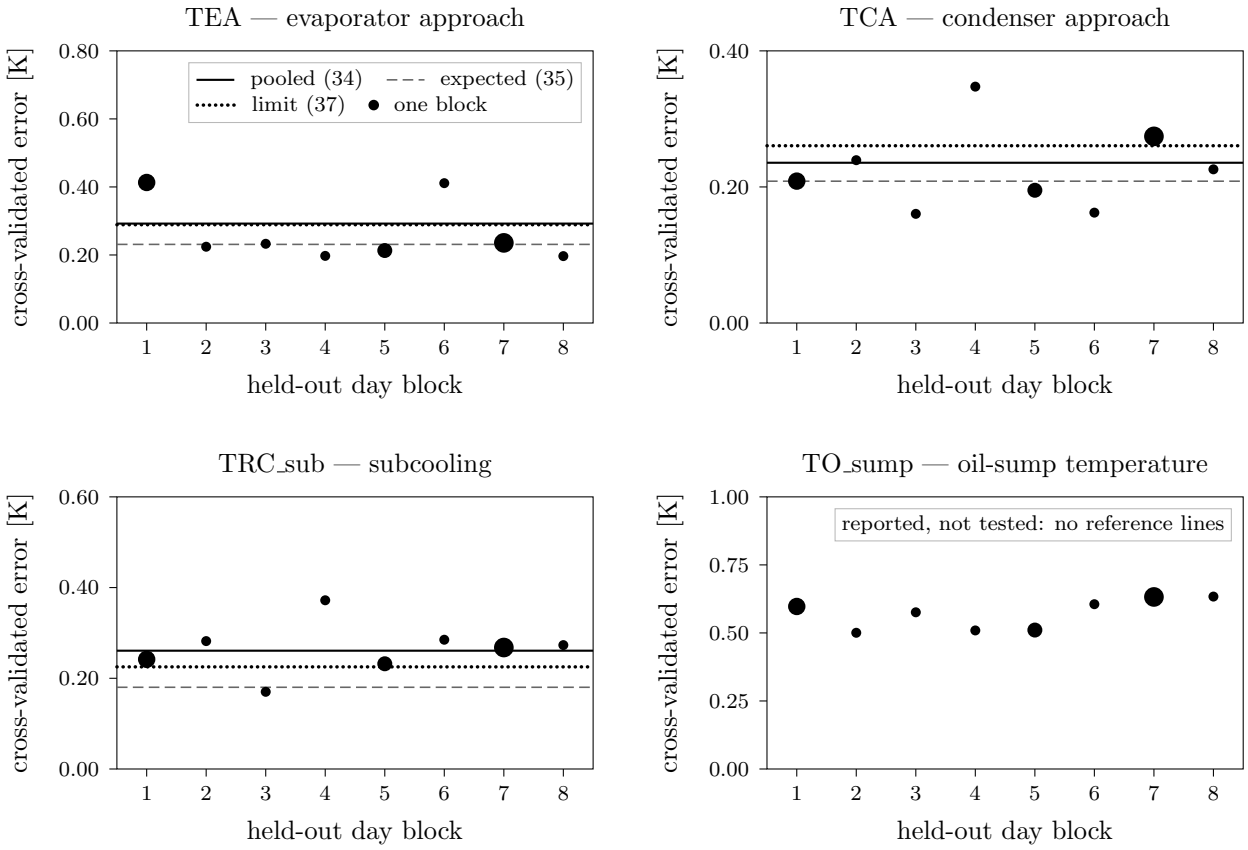


Figure 4: Cross-validated error of the refrigerant-side model, per channel and per held-out day block. Each dot is one block: the model is fitted on the other blocks and its predictions compared with that block's measurements, and the dot is the root-mean-square difference over the block's points; its area grows with the number of measured points in the block. In each tested panel the solid line is the pooled error CV_c of (34), the dashed line the error E_c a correct model would show given the measured scatter, from (35), and the dotted line the limit $1.25\,E_c$ of (37); the key in the first panel holds for all three tested panels. A channel fails when its solid line lies above its dotted line. The oil-sump panel draws no reference lines because that channel is reported and not tested: the protocol set no limit for it, and the panel says so.

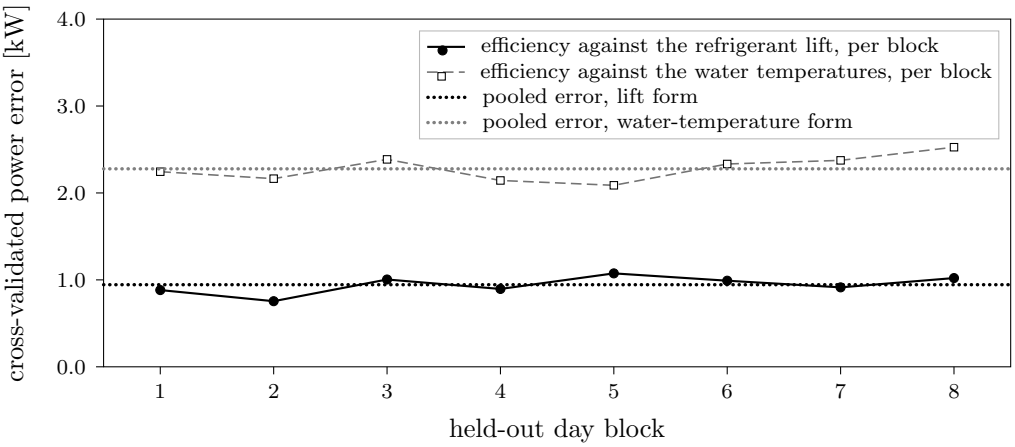


Figure 5: Cross-validated error of the two candidate compressor-power laws, per held-out day block. Each marker is the root-mean-square difference between the measured power and the power predicted by a law fitted without that block: filled circles on a solid line for the lift law of (2) and (3), whose efficiency is written against the refrigerant lift — the condenser shell's saturation temperature less the evaporating one — and open squares on a dashed line for the three-coefficient law of Gordon and Ng (1995) written in the water temperatures. The two dotted horizontal lines are the two forms' errors pooled over all blocks. The lift form lies below the water-temperature form in every block.

Table 9: Computations that could not be carried out, counted where they occurred. Every one is the same situation: at the sensor sets without refrigerant pressures there is no condensing temperature, so the shift a fault would make in the compressor power cannot be computed from the exchanger model. No fit and no verdict of this report depends on them.

Set	Fold [index]	Count [count]	Kind
L2	0	1	ValueError
L3	0	1	ValueError
L4	0	1	ValueError

shaded cells are pairs the measurements cannot separate. The table lists the same pairs with the number of standard deviations of reading the pair would require.

Recomputed on the normal runs, the families fall into three classes: {`cf`, `nc`, `rl`}, {`eo`}, {`ro`}. The protocol expected five, all of them single families. The disagreement is reported and governs nothing: the classes the protocol registered are the ones the companion report scores against. Two pairs drive the merge. Condenser fouling and non-condensable gas sit 0.867573 rad apart against a requirement of 1.28804 rad, and condenser fouling and a refrigerant leak 1.28291 rad apart against 1.28804 rad. All three degrade the condenser side in nearly the same direction at their mildest quantified severities. No family's signature lies inside the run-to-run spread of normal operation.

The merge is robust to the structural constants. Recomputing the classes at each end of each constant's range leaves it standing except at three ends — the condenser water-film share, the flooded area and the segregated share of the non-condensable gas — where it splits only partly. Whitening by the day-block scatter instead of the run-mean scatter gives {`cf`}, {`eo`, `nc`, `ro`}, {`rl`}, a different partition of the same three families rather than a separation of them, and the paired-contrast variant agrees. None of these governs any score.

The size of the non-condensable signature is worth stating, because it is the smallest of the

Table 10: Every pair of fault families at the laboratory instrumentation. *Angle* is the angle between the two signatures after dividing through by the measured scatter; *Threshold* is the angle below which the pair's own smallest quantified severities could be confused; *Magnitudes needed* is how many standard deviations of reading the pair would have to produce before that angle resolved them, against a run-to-run spread of 1 standard deviation. Pairs are ordered by name.

Family	Family	Angle [rad]	Threshold [rad]	Magnitudes needed [s.d.]	Verdict
cf	eo	1.32956	1.28804	1.62105	separated
cf	nc	0.867573	1.28804	2.3792	not separated
cf	rl	1.28291	1.28804	1.67123	not separated
cf	ro	1.85869	1.28804	1.24809	separated
nc	eo	1.2502	0.691163	1.70888	separated
rl	eo	1.7053	0.192202	1.32797	separated
rl	nc	2.15048	0.691163	1.13675	separated
rl	ro	3.14159	0.0737804	1	separated
ro	eo	1.43629	0.192202	1.51978	separated
ro	nc	0.991113	0.691163	2.10295	separated

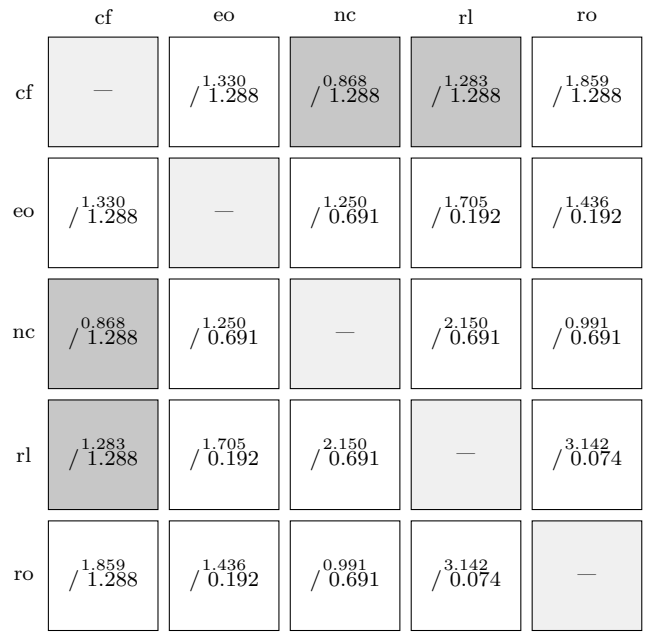
three. At the mildest quantified charge, nitrogen displacing 1 % of the condenser volume, the partial pressure it adds is 6088.02 Pa, which moves the condenser approach by 0.162575 K and the subcooling by 0.154896 K — both comparable with the day-to-day scatter of those channels in Table 4.

Detectability, which asks about noise rather than about another fault, is given per family and sensor set in Figure 7 and Table 11. Each panel of the figure holds one sensor set; for each family the range of quantified severities is a gray bar and the smallest detectable severity of (41) a vertical rule, so a rule left of its bar means every quantified case is within reach. With the laboratory instrumentation, condenser fouling is expected to be resolvable from 7.46839 % of tubes plugged against a mildest quantified case of 6 %, so that case is out of reach; non-condensable gas needs 1.05687 % of the condenser volume against a mildest quantified 1 %, which is marginal. Both flow faults are resolvable far below their mildest quantified severities wherever a flow meter survives. At the three lower sensor sets the following families are beyond reach even at their deepest quantified severity: excess oil, refrigerant leak. Every family reaches at least one channel at every sensor set, so nothing is invisible in the strict sense; what the lower sets take away is resolution, not reach.

The recomputed classes are set beside the registered ones because the protocol fixed an expectation that normal operation can test. The difference — condenser-side faults merged where single families were expected — means that the measured scatter of this machine is large, relative to the mildest quantified severities, against the differences between those faults' effects on the four channels. The comparison across sensor sets in the detectability results makes the same point for noise alone. Together they demonstrate that identifiability and detectability are properties of the model and the instrumentation jointly, and that both can be computed from normal operation before any fault is scored.

4.4 Predicted false alarms

The rate at which the model, shown a simulated normal run drawn from its own fitted description, answers something other than normal is plotted per sensor set in Figure 8 and counted in Table 12. The bars are the pooled rate of (42) per sensor set, the dots the same rate within each held-out block, and the table names the wrong answers that occur most often. With the laboratory



Classes the protocol expected:
{`cf`}, {`eo`}, {`nc`}, {`rl`}, {`ro`}

Classes normal operation supports:
{`cf`, `nc`, `rl`}, {`eo`}, {`ro`}

A shaded cell is a pair the measurements cannot separate: its angle is below the angle the pair's own magnitudes would need. Angles are in radians; the second number in each cell is the threshold.

Figure 6: Angles between the fault signatures at the laboratory instrumentation. A fault's signature is the direction in which it moves the four measured refrigerant-side channels, divided through by the measured scatter of those channels, so that the length of a signature is the number of standard deviations of reading the fault would produce. Two faults are separable when the angle between their signatures is large enough that their smallest quantified severities cannot be confused; the threshold beside each angle is that requirement, computed from the two signatures' own lengths. Shaded cells are pairs the measurements cannot separate. The diagonal is empty by construction.

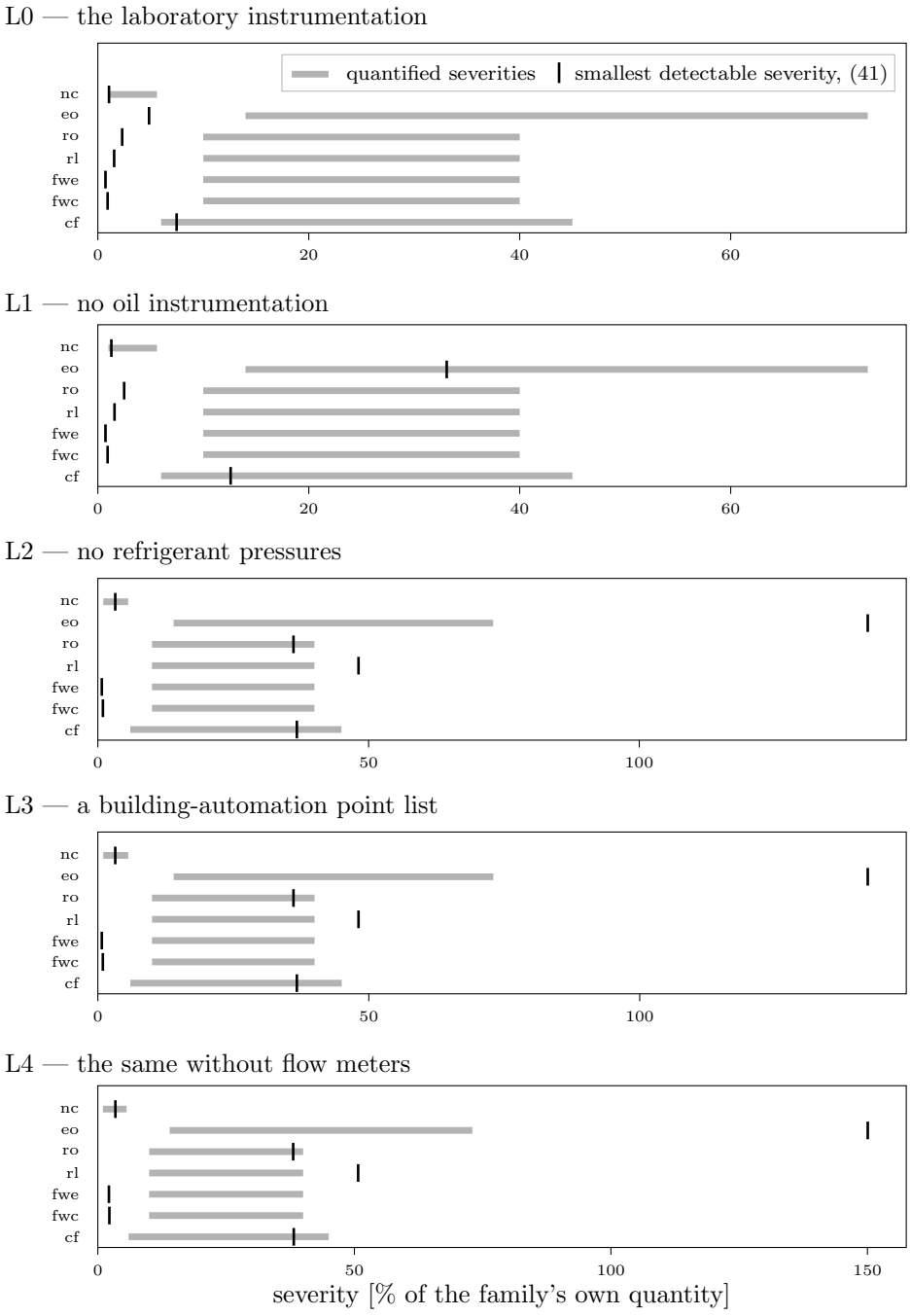


Figure 7: Smallest detectable severity per fault family, at each sensor set, one panel per set from the laboratory instrumentation (top) to the building-automation list without flow meters (bottom). The severity unit is the family's own: percent of condenser tubes plugged, percent of water-flow reduction, percent of refrigerant charge removed or added, percent of excess oil, and percent of the condenser volume occupied by non-condensable gas. The gray bar spans the severities the benchmark quantifies; the vertical rule is the smallest severity of (41) that the fitted model and the measured scatter would let a test of the stated size and power resolve, and the key in the top panel holds for all five. A rule left of its bar means every quantified severity is expected to be detectable. Families whose rule lies beyond the plotted range cannot be detected at any quantified severity from that sensor set.

Table 11: Detectability per fault family and sensor set. *Lowest* and *Highest* bound the severities the benchmark quantifies, in the family's own unit; *Detectable from* is the smallest severity a test of size 0.05 and power 0.8 would resolve given the fitted model and the measured scatter; *Ratio* is that severity divided by the lowest quantified one, so a value above one means the mildest quantified case is out of reach. Families are in the order the protocol fixed.

Set	Family	Lowest [%]	Highest [%]	Detectable from [%]	Ratio [–]	At the lowest severity
L0	cf	6	45	7.46839	1.25745	expected undetectable
L0	fwc	10	40	0.924415	0.0881463	expected detectable
L0	fwe	10	40	0.723769	0.0689443	expected detectable
L0	rl	10	40	1.54448	0.154448	expected detectable
L0	ro	10	40	2.30889	0.230889	expected detectable
L0	eo	14	73	4.85708	0.346934	expected detectable
L0	nc	1	5.6	1.05687	1.05687	expected undetectable
L1	cf	6	45	12.5881	2.19844	expected undetectable
L1	fwc	10	40	0.930102	0.0886912	expected detectable
L1	fwe	10	40	0.727177	0.0692702	expected detectable
L1	rl	10	40	1.57916	0.157916	expected detectable
L1	ro	10	40	2.49527	0.249527	expected detectable
L1	eo	14	73	33.0768	2.36263	expected undetectable
L1	nc	1	5.6	1.28456	1.28456	expected undetectable
L2	cf	6	45	36.754	7.8611	expected undetectable
L2	fwc	10	40	0.932722	0.0889422	expected detectable
L2	fwe	10	40	0.728521	0.0693986	expected detectable
L2	rl	10	40	48.1477	4.81477	expected undetectable
L2	ro	10	40	36.131	3.6131	expected undetectable
L2	eo	14	73	142.199	10.1571	expected undetectable
L2	nc	1	5.6	3.21967	3.21967	expected undetectable
L3	cf	6	45	36.7345	7.85544	expected undetectable
L3	fwc	10	40	0.927641	0.0884554	expected detectable
L3	fwe	10	40	0.728944	0.0694391	expected detectable
L3	rl	10	40	48.1121	4.81121	expected undetectable
L3	ro	10	40	36.1052	3.61052	expected undetectable
L3	eo	14	73	142.086	10.149	expected undetectable
L3	nc	1	5.6	3.21688	3.21688	expected undetectable
L4	cf	6	45	38.1925	8.28439	expected undetectable
L4	fwc	10	40	2.24812	0.215809	expected detectable
L4	fwe	10	40	2.15972	0.20723	expected detectable
L4	rl	10	40	50.7346	5.07346	expected undetectable
L4	ro	10	40	38.0778	3.80778	expected undetectable
L4	eo	14	73	150.059	10.7185	expected undetectable
L4	nc	1	5.6	3.4242	3.4242	expected undetectable

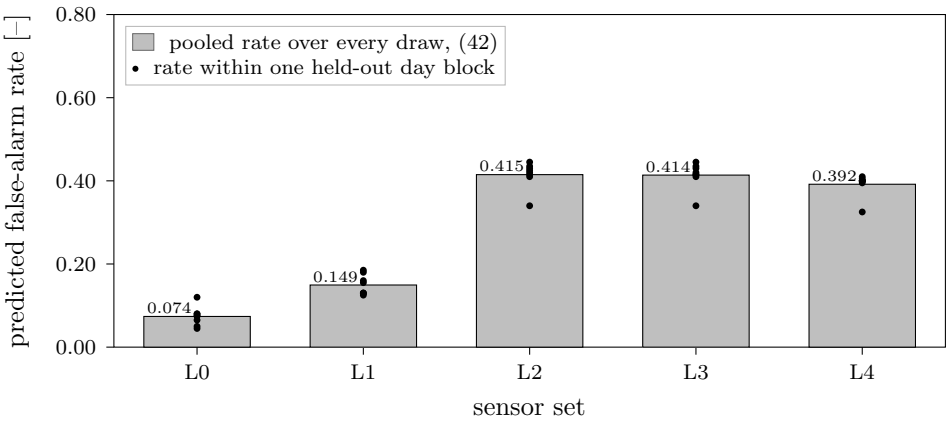


Figure 8: Predicted rate at which the model calls a normal run faulty, per sensor set, from 1,600 simulated normal runs at each set. Each draw takes a normal day block's own fitted description, draws a fresh set of readings from it with no fault present, and asks the model which of its eight hypotheses is most probable; a draw counts as a false alarm when the answer is not normal. The bars are the pooled rate of (42), printed above each bar, and the dots the same rate within each held-out day block. The rate is a property of the model and its priors, computed before any measured fault run was read.

Table 12: Predicted false alarms per sensor set: how often the model, shown a simulated normal run drawn from its own fitted description of normal operation, answers something other than normal. The last column names the wrong answers that occur most often, with their counts. The rate is a prediction made before any measured fault run was read, and the companion report scores the realized rate against it.

Set	Draws [count]	False alarms [count]	Rate [–]	Most frequent wrong answers [count each]
L0	1,600	118	0.07375	nc 58, cf 46, eo 7, rl 6
L1	1,600	239	0.149375	cf 84, nc 79, eo 67, rl 7
L2	1,600	664	0.415	rl 344, ro 214, nc 92, cf 14
L3	1,600	662	0.41375	rl 343, ro 213, nc 92, cf 14
L4	1,600	627	0.391875	rl 276, ro 233, nc 109, cf 8

instrumentation the rate is 0.07375, 118 of 1,600 draws; removing the oil instrumentation raises it to 0.149375; removing the refrigerant pressures raises it to 0.415, and it stays near that value at the two sets below, 0.41375 and 0.391875. The wrong answers change character with the sets: at the two upper sets they are mostly non-condensable gas and condenser fouling, the two families whose signatures are smallest relative to the scatter; at the three lower sets they are mostly the refrigerant faults, which the remaining channels cannot constrain at all and which the prior therefore carries. A rate near 0.4 is not a property of the inference but of the question asked of it: with only the compressor power and the water temperatures left, several hypotheses explain an untroubled run about as well as normal does, and a most-probable-answer rule under a uniform prior over eight hypotheses will pick one of them often.

The rates are compared across the sensor sets because the sets stand for the instrumentation a real building carries. The rise when the refrigerant pressures are removed means that, without them, normal operation is no longer distinguishable from several mild faults at the scatter this machine shows. The comparison demonstrates that the false-alarm rate of a model-based diagnosis is a property of the model and the instrumentation that can be predicted before a single fault run is read.

Table 13: Validation on simulated plants. *Identified* is the fraction of simulated faults whose most probable hypothesis is the simulated family; *False alarms* counts simulated normal plants answered as faulty; *Coverage* is the fraction of stated magnitude intervals containing the simulated magnitude, against a nominal 0.95; *Missed* counts simulated faults above the model's own stated detection threshold that it nonetheless missed, which should be zero. *Confidence* is the mean probability the model attaches to its own answer and *Accuracy* how often that answer is right, with an exact binomial interval: a calibrated model matches the two. Every plant in this table is simulated.

Scatter	Set	Identified [–]	False alarms [count]	Coverage [–]	Missed [count]	Confidence [–]	Accuracy [–]
measured	L0	0.935484	0 of 16	0.935484	0	0.877921	0.8375 [0.738163, 0.910538]
measured	L1	0.83871	0 of 16	0.943548	0	0.811303	0.7875 [0.681714, 0.871064]
measured	L2	0.387097	7 of 16	0.951613	0	0.481479	0.5 [0.386048, 0.613952]
measured	L3	0.387097	7 of 16	0.951613	0	0.481479	0.5 [0.386048, 0.613952]
measured	L4	0.41129	8 of 16	0.959677	0	0.485976	0.55 [0.434674, 0.661514]
quiet	L0	0.951613	2 of 16	0.927419	0	0.891244	0.8875 [0.797153, 0.947248]
quiet	L1	0.846774	0 of 16	0.919355	0	0.845476	0.8 [0.695631, 0.881141]
quiet	L2	0.516129	0 of 16	0.967742	0	0.60216	0.6125 [0.496976, 0.719435]
quiet	L3	0.516129	0 of 16	0.967742	0	0.60216	0.6125 [0.496976, 0.719435]
quiet	L4	0.556452	3 of 16	0.967742	0	0.609172	0.6 [0.484377, 0.707991]

4.5 Validation on simulated plants

Everything in this subsection is computed on plants simulated from the model, never on measurements; its purpose is to separate a failure of the computation from a failure of the model and from a limit of what the measurements carry. The headline quantities are set out in Table 13 at the scatter measured on the real normal runs and again on a quieter plant, one row per scatter and sensor set, and Figure 9 breaks the identification rate down by family.

At the measured scatter the identification rate of (43), pooled over families, is 0.935484 with the laboratory instrumentation and 0.83871 without the oil instrumentation, falling to 0.387097–0.41129 at the three lower sets, which is the same collapse the detectability analysis of Section 4.3 predicts and for the same reason. Simulated normal plants are answered as faulty 0 times in 16 at the laboratory instrumentation. The coverage of (44), the fraction of magnitude intervals containing the simulated magnitude, is 0.935484–0.959677 across the sets against a nominal 0.95, and zero simulated faults above the model's own stated detection threshold were missed, which is the outcome an implementation without a defect gives. The calibration check agrees: the mean stated confidence of (45) tracks the accuracy of (46) at every set — at L2, 0.481479 against 0.5 with an interval of [0.386048, 0.613952] — and its intervals cover at 0.892308–0.955882, every one of them consistent with the nominal 0.95. Whatever the model gets wrong on real measurements, it does not get it wrong because the fitting or the evidence computation is broken.

Figure 10 and Table 14 ask a different question: what happens when the simulated plant does not follow the model's own mechanism. Each alternative mechanism is paired with a matched control — the same draws with the model's mechanism in place — so that the comparison isolates the mismatch rather than the difficulty of the case.

What moves most is the magnitude interval. Its largest loss is under condenser pool geometry 0.2 at L1, where coverage falls to 0.375 from 0.916667 in the control; the same happens, less severely, under the valve's authority and under the distribution of non-condensable gas. Each of these is a mechanism the model could carry as a further structural constant and does not. Two effects are smaller but real and are not to be passed over. The label moves furthest under non-condensable segregation 1 at L1, where the simulated family is recovered in 0.416667 of draws

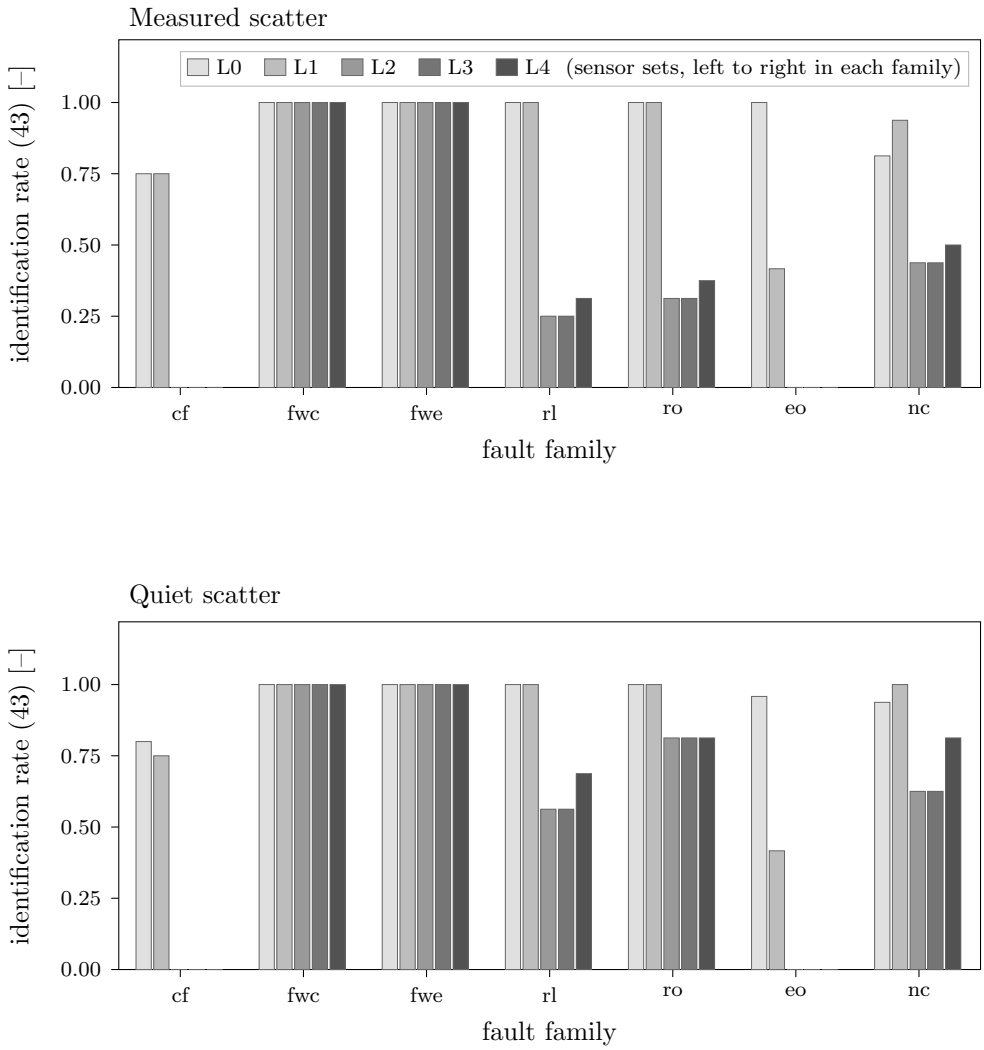


Figure 9: Identification on simulated plants. For every family and sensor set, a plant with that fault at a quantified severity is simulated, its readings are drawn with the stated scatter, and the model is asked which hypothesis is most probable; the bar is the identification rate of (43), the fraction of draws whose answer is the family that was simulated. Within each family the five bars are the five sensor sets, from L0 (lightest, left) to L4 (darkest, right), as the key in the upper panel says. The upper panel uses the scatter measured on the real normal runs, the lower panel a quieter plant, which separates what the inference cannot do from what the noise does not permit. Every plant here is simulated: no measured fault run enters this figure.

against 1 in the control. And a day-block conductance drift raises the false-alarm rate: under condenser conductance drift 0.02 at L0 untroubled plants are answered as faulty at 0.3125 against $< 10^{-9}$ in the control, because a slow drift of a conductance is the shape a mild condenser fouling has. A separate program repeats that question on 48 simulated normal plants, with the drift placed in the training block, in the diagnosed run, or in both: the model's own plant 6 of 48; two-point drift in the baseline block and in the run 5 of 48; the same on the condenser alone 10 of 48; condenser drift in the run only 9 of 48; two-point drift in the run only 4 of 48. Tables 15 and 16 classify every weakness the three programs found, separating the mechanisms a further build could represent from the quantities no implementation recovers.

Two comparisons close this subsection. The measured and the quieter scatter are compared because their difference separates what the inference cannot do from what the noise does not permit: where identification recovers on the quieter plant, its loss at the measured scatter is a limit of the measurements and not of the inference. Each alternative mechanism is compared with its matched control because the control holds the case and the draws fixed and changes only the mechanism, so the difference is the cost of a representation the model lacks; it falls mainly on the magnitude interval of (44) rather than on the label. Together the comparisons demonstrate that validation on simulated plants attributes each weakness to the computation, to the representation or to the data before any real fault is diagnosed.

4.6 Consequences of the failed test for the confirmatory comparison

Because the acceptance test failed, the fallback the protocol fixed governs what follows: the confirmatory comparison of the companion report runs not on all 39 scored runs but on the restricted population of 17. Table 17 states what that costs, as a property of the design and not as a result. Two methods are compared run by run and only the runs on which exactly one of them is right carry information; at a split in which one method wins three runs in ten and the other none, the paired test (McNemar 1947) would reject at the five-percent level with probability 0.989111 on the full population and 0.403181 on the restricted one, and 0.10464 at the level a correction for four comparisons (Holm 1979) imposes. The confirmatory comparison is therefore underpowered by construction, and the companion report says so before it reports an outcome.

5 Discussion

The central result is negative and it is specific. The refrigerant-side model does not reproduce normal operation to the strictness the protocol demanded, on two of three channels, and the shortfall is day-to-day: the evaporator approach's error concentrates in two of eight day blocks. What the conductance laws of (1) carry is the dependence of exchanger conductance on load and water temperature; what they do not carry is a shift of the refrigerant side that persists through a day of testing and then changes. That is a finding about the model's adequacy at the registered strictness, not a defect of the inference: the validation of Section 4.5 recovers what it should on simulated plants and its stated confidences match its accuracy, which is what an implementation failure would have broken. It is also not a large absolute error — the unexplained part is 0.178767 K on a channel whose day-to-day scatter is of the same order — which is precisely why a limit fixed in advance was needed to decide it.

Two features of the test deserve to be weighed against the verdict rather than hidden. The reference against which the error is judged was chosen after the point-level ratios had been seen, and the numerators of two channels were known when the limit was set; a test constructed that way can be accused of being aimed. Against that stand the stricter form, whose numerators were

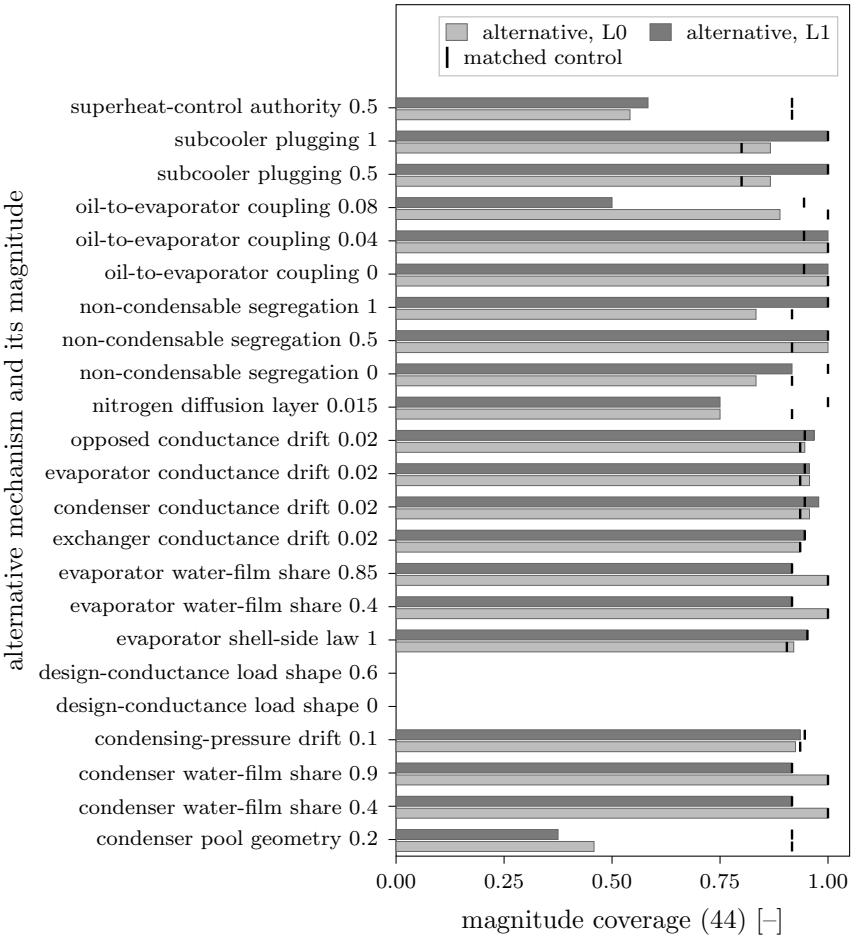


Figure 10: Coverage of the magnitude interval when the simulated plant does not follow the model's own mechanism, at the scatter measured on the real normal runs. Each pair of bars is one alternative mechanism at the magnitude printed beside it, the lower bar at L0 and the upper at L1; a bar is the coverage of (44), the fraction of draws whose stated interval contains the simulated magnitude, and the rule across it is the same quantity in a matched control where the plant does follow the model. A bar well short of its rule marks a mechanism the model represents too narrowly; a bar at its rule marks one the measurements could not have distinguished in any case. A mechanism whose plant reaches no case of the catalog at a sensor set draws no bar there. Every plant here is simulated.

Table 14: Simulated plants that do not follow the model's own mechanism, at the scatter measured on the real normal runs. Each row pairs the quantity measured when the plant follows the alternative mechanism with the same quantity in a matched control where it does not, written as *alternative / control*; *Draws* counts the faulty draws and then the normal ones. A mechanism is exercised only on the fault cases it can reach, so a mechanism that reaches no case at a sensor set draws nothing there and a mechanism that alters no normal plant is given no normal draws; a dash marks a quantity those zero draws leave undefined. Every plant in this table is simulated.

Alternative mechanism	Set	Draws [count]	Identified [–]	Coverage [–]	False alarms [–]
condenser pool geometry 0.2	L0	24 / 0	1 / 1	0.458333 / 0.916667	— / —
condenser pool geometry 0.2	L1	24 / 0	1 / 1	0.375 / 0.916667	— / —
condenser water-film share 0.4	L0	12 / 0	1 / 1	1 / 1	— / —
condenser water-film share 0.4	L1	12 / 0	1 / 1	0.916667 / 0.916667	— / —
condenser water-film share 0.9	L0	12 / 0	1 / 1	1 / 1	— / —
condenser water-film share 0.9	L1	12 / 0	1 / 1	0.916667 / 0.916667	— / —
condensing-pressure drift 0.1	L0	93 / 16	0.913978 / 0.913978	0.924731 / 0.935484	$< 10^{-9}$ / $< 10^{-9}$
condensing-pressure drift 0.1	L1	93 / 16	0.870968 / 0.870968	0.935484 / 0.946237	0.1875 / $< 10^{-9}$
design-conductance load shape 0	L0	0 / 0	— / —	— / —	— / —
design-conductance load shape 0	L1	0 / 0	— / —	— / —	— / —
design-conductance load shape 0.6	L0	0 / 0	— / —	— / —	— / —
design-conductance load shape 0.6	L1	0 / 0	— / —	— / —	— / —
evaporator shell-side law 1	L0	63 / 0	0.873016 / 0.873016	0.920635 / 0.904762	— / —
evaporator shell-side law 1	L1	63 / 0	0.936508 / 0.936508	0.952381 / 0.952381	— / —
evaporator water-film share 0.4	L0	12 / 0	1 / 1	1 / 1	— / —
evaporator water-film share 0.4	L1	12 / 0	1 / 1	0.916667 / 0.916667	— / —
evaporator water-film share 0.85	L0	12 / 0	1 / 1	1 / 1	— / —
evaporator water-film share 0.85	L1	12 / 0	1 / 1	0.916667 / 0.916667	— / —
exchanger conductance drift 0.02	L0	93 / 16	0.935484 / 0.913978	0.935484 / 0.935484	0.125 / $< 10^{-9}$
exchanger conductance drift 0.02	L1	93 / 16	0.817204 / 0.870968	0.946237 / 0.946237	0.1875 / $< 10^{-9}$
condenser conductance drift 0.02	L0	93 / 16	0.967742 / 0.913978	0.956989 / 0.935484	0.3125 / $< 10^{-9}$
condenser conductance drift 0.02	L1	93 / 16	0.817204 / 0.870968	0.978495 / 0.946237	0.125 / $< 10^{-9}$
evaporator conductance drift 0.02	L0	93 / 16	0.946237 / 0.913978	0.956989 / 0.935484	0.0625 / $< 10^{-9}$
evaporator conductance drift 0.02	L1	93 / 16	0.827957 / 0.870968	0.956989 / 0.946237	0.125 / $< 10^{-9}$
opposed conductance drift 0.02	L0	93 / 16	0.924731 / 0.913978	0.946237 / 0.935484	$< 10^{-9}$ / $< 10^{-9}$
opposed conductance drift 0.02	L1	93 / 16	0.83871 / 0.870968	0.967742 / 0.946237	0.125 / $< 10^{-9}$
nitrogen diffusion layer 0.015	L0	12 / 0	0.75 / 0.75	0.75 / 0.916667	— / —
nitrogen diffusion layer 0.015	L1	12 / 0	1 / 1	0.75 / 1	— / —
non-condensable segregation 0	L0	12 / 0	1 / 0.75	0.833333 / 0.916667	— / —
non-condensable segregation 0	L1	12 / 0	1 / 1	0.916667 / 1	— / —
non-condensable segregation 0.5	L0	12 / 0	0.916667 / 0.75	1 / 0.916667	— / —
non-condensable segregation 0.5	L1	12 / 0	1 / 1	1 / 1	— / —
non-condensable segregation 1	L0	12 / 0	0.666667 / 0.75	0.833333 / 0.916667	— / —
non-condensable segregation 1	L1	12 / 0	0.416667 / 1	1 / 1	— / —
oil-to-evaporator coupling 0	L0	18 / 0	1 / 1	1 / 1	— / —
oil-to-evaporator coupling 0	L1	18 / 0	0.0555556 / 0.555556	1 / 0.944444	— / —
oil-to-evaporator coupling 0.04	L0	18 / 0	1 / 1	1 / 1	— / —
oil-to-evaporator coupling 0.04	L1	18 / 0	0.944444 / 0.555556	1 / 0.944444	— / —
oil-to-evaporator coupling 0.08	L0	18 / 0	1 / 1	0.888889 / 1	— / —
oil-to-evaporator coupling 0.08	L1	18 / 0	1 / 0.555556	0.5 / 0.944444	— / —
subcooler plugging 0.5	L0	15 / 0	0.666667 / 0.666667	0.866667 / 0.8	— / —
subcooler plugging 0.5	L1	15 / 0	0.733333 / 0.733333	1 / 1	— / —
subcooler plugging 1	L0	15 / 0	0.666667 / 0.666667	0.866667 / 0.8	— / —
subcooler plugging 1	L1	15 / 0	0.733333 / 0.733333	1 / 1	— / —
superheat-control authority 0.5	L0	24 / 0	1 / 1	0.541667 / 0.916667	— / —
superheat-control authority 0.5	L1	24 / 0	1 / 1	0.583333 / 0.916667	— / —

Table 15: The weaknesses the validation classified as mechanisms the model does not carry, in the protocol's own record of them. Each is a mechanism a further build could represent as one more structural constant with its own prior; the alternative-mechanism plants of Table 14 show what each costs, and in every case what it costs is the magnitude interval and not the label. The remaining weaknesses are in Table 16.

Id	Weakness	Kind	What settles the classification
A1c	excess-oil churning power not modeled	representation (below the power scatter)	at eo14 the registered UA_oil range gives 0.04-0.18 kW, inside the power channel's run scatter; not built
R1	the charge map without the condenser pool's segment geometry	representation (magnitude only)	on a synthetic truth with a bundle clearance of 0.20 (the end of the registered range) leak and overcharge are still identified at every level, but rl / ro magnitude coverage falls from 0.92 to 0.46 (L0) and 0.38 (L1) and the overcharge magnitude is overstated by about 28 percent; the registered log-sd widening cannot cover a shift. A further build: integrate the clearance over 0-0.20 as a structural constant
R2	the charge map without superheat-valve control	representation (magnitude only)	at half authority rl / ro labels are unchanged and magnitude coverage falls from 0.92 to 0.54 (L0) and 0.58 (L1), to 0.33-0.42 for the leak alone. A further build: integrate the authority over 0-0.5
R3	the nitrogen diffusion layer registered at zero	representation (magnitude only)	at 1.5 percent of UA_c per volume percent the nc label is unchanged and magnitude coverage falls from 0.92-1.00 to 0.75 (12 draws, not significant at realistic noise; quiet L1 0.67 against 1.00). A further build: a second axis of the share quadrature

Table 16: The weaknesses the validation classified as limits of what the measurements carry, together with the findings that show no fault, in the protocol's own record of them. A *data limit* is a quantity no implementation recovers, because the measurements do not contain it; the detectability analysis and the quieter simulated plant are what settle that. No weakness in either table is a defect of the fitting or of the evidence computation, because the self-recovery and calibration results rule that out.

Id	Weakness	Kind	What settles the classification
D1	the eo label at L1 depends on the unknowable coupling	data limit	k 0: identification 0.06 (control 0.56); k 0.04: 0.94; k 0.08: 1.00 with coverage 0.50 and log bias +0.37; every k is inside the registered prior
D2	eo magnitude at the coupling's support edge at L2-L4	data limit	k 0.08: coverage 0.78-0.89 (control 0.94-1.00); labels 0.00 at every k, as W4
D3	nc lost when the nitrogen is fully segregated	data limit	s = 1: L1 0.42 (control 1.00), L4 0.08 (0.50), L2 / L3 0.25 (0.50), L0 0.67 (0.75); a fully segregated pocket is a condenser-conductance loss, the cf direction, with no Dalton offset to separate it
D4	condenser-only day-block drift at L0	data limit	false alarms 10/48 against the model's own 6/48 (not significant); the drift is the model's own state and cf-shaped by physics
W1	cf6 / cf12 called normal at L0, L1 (identification 0.25-0.50)	data limit	the smallest condenser fouling sits inside day-block conductance drift, which is cf-shaped by physics; the model's own detectability predictor calls cf6 undetectable
W2	nc1 called normal at L0 (identification 0.50)	data limit	nc1's Dalton offset (about 0.28 K) sits inside the TCA / TRC_sub run scatter
W3	eo missed at L1 (0.42 overall; eo14 0.00, called normal)	data limit	without the sump sensor eo is read only through an oil-to-evaporator coupling whose prior has half its mass at zero
W4	at L2-L4 cf and eo are not identified and rl, ro, nc are weak (0.25-0.50)	data limit	at the benchmark's power scatter the levels without refrigerant pressures identify only the flow faults; the quiet plant recovers rl, ro and nc, so the cause is the power scatter
W5	false alarms on the model's own normals at L2-L4: 7/16, 7/16, 8/16	data limit	at the levels without refrigerant pressures the model's own false-alarm rate on normals is about 0.45 under the registered uniform model prior, because small-magnitude rl is near-aliased with normal on power alone; this report's predictor reports it per fold
W6	rows predicted undetectable detected 4/4 at L2-L4 (cf12-cf45, ro30-ro40, nc1-nc2.4)	finding, non-failing direction	the predictor measures detectability against normal while a wrong family's detection counts as detection
W7	calibration magnitude coverage at L0 0.89 [0.79, 0.96]	open, no fault shown	passes; resolving 0.89 against 0.95 needs about 400 draws

Table 17: What the confirmatory comparison of the companion report can resolve, as a property of the design and not as a result. Two methods are compared run by run, and only the runs on which exactly one of them is right carry information. The table gives the probability that the paired test rejects, for each split of those runs and each significance level, on the full scored population and on the restricted population the protocol falls back to. Rates printed as bounds are zero to round-off.

Population	Runs [count]	Level [–]	Rate A only [–]	Rate B only [–]	Probability of a verdict [–]
restricted	17	0.0125	0.3	$< 10^{-9}$	0.10464
restricted	17	0.0125	0.25	0.05	0.0231984
restricted	17	0.0125	0.2	$< 10^{-9}$	0.0109343
restricted	17	0.0125	0.15	0.05	0.00103106
restricted	17	0.05	0.3	$< 10^{-9}$	0.403181
restricted	17	0.05	0.25	0.05	0.129344
restricted	17	0.05	0.2	$< 10^{-9}$	0.105701
restricted	17	0.05	0.15	0.05	0.0172591
full	39	0.0125	0.3	$< 10^{-9}$	0.933572
full	39	0.0125	0.25	0.05	0.318734
full	39	0.0125	0.2	$< 10^{-9}$	0.531658
full	39	0.0125	0.15	0.05	0.0554344
full	39	0.05	0.3	$< 10^{-9}$	0.989111
full	39	0.05	0.25	0.05	0.552553
full	39	0.05	0.2	$< 10^{-9}$	0.819987
full	39	0.05	0.15	0.05	0.16522

not known and which fails on the same two channels, and the third channel, whose numerator was not known and which fails by the largest margin of the three. The feature sets were also chosen on these same runs, which makes the measured error too small rather than too large; a test that fails under a generous measurement fails.

Identifiability on the real normal runs is coarser than the protocol expected. Condenser fouling, non-condensable gas and a refrigerant leak share a direction: all three reduce the effective condenser conductance at their mildest quantified severities, fouling by removing area, nitrogen by blanketing part of the bundle, and a leak by lowering the liquid pool and with it the wetted area. Separating them needs either a larger severity or a measurement the benchmark does not carry, such as a direct measure of refrigerant charge or a second temperature across the subcooler. This changes no registered score, and it does predict where the companion report will find its confusions.

The false-alarm rate near 0.4 at the three lower sensor sets is a limit of what those measurements carry rather than a modeling failure, and it is worth saying why the number is not smaller. The model answers with its most probable hypothesis under a uniform prior over eight, and with only the compressor power and the water temperatures left, several fault hypotheses explain an untroubled run about as well as normal does. An operator would not run such a rule; a decision threshold, or a prior weighted by how often faults actually occur, would trade those false alarms for missed mild faults. Reporting the rate under the flat rule is the honest baseline, and the companion report measures the realized rate against it. A related discrepancy sits underneath: the normal runs imply a full-load condenser approach several times the machine's design value, which the protocol discloses; the design-approach form of (2) and (3) inherits that error, which is visible in its cross-validated error above.

The alternative-mechanism program points at what to build next rather than at what to disclaim. The mechanisms whose magnitude coverage falls — the condenser pool's geometry, the expansion valve's authority, the distribution of non-condensable gas — are all representable; each

Table 18: The answers of this report, one row per question in the order of the Conclusion. *Answer* is one of yes, conditional, no or confounded with its qualifier; *Why* gives the number that decides it and the relation that computed it; *What would change it* names the measurement, action or modeling change the report identifies, or says that it identifies none. Every number is restated from the Results; no verdict here is new.

Question	Answer	Why	What would change it
Does the refrigerant-side model reproduce normal operation?	*no*, on two of three channels	R_c of (36) is 1.26425 and 1.4481 against $\kappa = 1.25$	a law that carries a shift of the refrigerant side persisting through a day of testing
Is the lift form of the power law the better one?	*yes*, in eight of eight blocks	pooled error 0.944495 against 2.2775 kW by (34)	losing the refrigerant pressures; the rebuilt lift gives 2.28919 kW
Can condenser fouling, non-condensable gas and a leak be told apart?	*no*, at the mildest quantified severities	angle 0.867573 against 1.28804 rad by (40)	a larger severity, a direct measure of charge, or a second subcooler temperature
Is the mildest quantified condenser fouling detectable at L0?	*no*	smallest detectable 7.46839 % against 6 % by (41)	nothing identified
Does the false-alarm rate stay at its L0 level without the refrigerant pressures?	*no*, it rises several-fold	rate 0.07375 at L0, 0.415 at L2 by (42)	a decision threshold, or a prior weighted by how often faults occur
Is the inference itself sound?	*yes*, on simulated plants	confidence tracks accuracy and the intervals cover at their nominal level, (44)–(46)	nothing identified
Can the confirmatory comparison resolve a three-in-ten split?	*conditional*, on the population	probability 0.403181 on 17 runs, 0.989111 on 39	a passing acceptance test, which restores the full population

would enter as a further structural constant with its own prior, integrated out by (7) exactly as the segregated share of the gas and the oil coupling of Section 3.6 already are, and Table 15 states what each would buy. The cases where no way of writing the model recovers the magnitude are different in kind and are marked as limits of the measurements in Table 16.

The limits of this report are plain. It concerns one machine, in a laboratory, with seventeen normal runs over eight day blocks of testing; the day-to-day component that the acceptance test caught is estimated from those same blocks and is the least well determined quantity in the analysis. The validation is on plants simulated from the model itself, so it bounds what the inference can do and not what the machine will do. And nothing here is evidence about faults: no fault run was read.

The verdicts are gathered in Table 18, one row per question a user of this machine's record would ask of it, with the number that decides each and what would change it.

6 Conclusion

This report asked, of a physics-structured Bayesian diagnosis model for a water-cooled centrifugal chiller and of seventeen normal runs of a laboratory benchmark, four questions, and fixed the criteria for answering them before computing any of them.

Whether the refrigerant-side model reproduces normal operation: it does not, at the strictness set in advance. Its cross-validated error is 1.26425 times the expected error on the evaporator approach and 1.4481 on the subcooling, against a limit of 1.25; the condenser approach passes at 1.1302. The stricter form of the same test fails on the same channels, and the shortfall is a day-to-day shift of the refrigerant side that the conductance laws do not represent.

Which form of the compressor-power law the measurements select: the one written against the refrigerant lift, which gives 0.944495 kW of cross-validated error against 2.2775 kW for the water-temperature form, and is better in every held-out day block.

Which faults the instrumentation can tell apart: fewer than expected. Condenser fouling, non-condensable gas and a refrigerant leak form one class at the mildest quantified severities, against the five single families the protocol registered, and the three lower sensor sets cannot resolve the refrigerant faults at any quantified severity.

How often the model would call an untroubled machine faulty, and whether the inference itself is sound: 0.07375 with the laboratory instrumentation and about 0.4 without the refrigerant pressures; and the inference recovers what it should on simulated plants, with stated confidences that match its accuracy and intervals that cover at their nominal rate. The model's weaknesses are therefore ones of representation and of what the measurements carry, not of the computation.

Because the acceptance test failed, the fallback fixed in advance governs what comes next: the confirmatory comparison of the companion report runs on 17 runs rather than 39, with the loss of resolving power stated in Table 17, and the physics-based claims are restricted to the classes the machine demonstrably expresses.

A Reproduction

The computation on normal operation is one command, run once,

```
python run.py stage1 --jobs 5
```

under the environment settings `OPENBLAS_CORETYPE=Haswell`, `OPENBLAS_NUM_THREADS=1`, which took 2.36076 hours on the machine described below. It reads the committed measurement extract

and the protocol file and writes the results file this report is built from; the report is regenerated from that file by a second command and contains no hand-entered number. The seeds are fixed in the protocol: the draws of the false-alarm prediction and of the detection predictions use 1043, and the validation programs use the seed the protocol records for them. 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 on normal operation was repeated under a second kernel and the two runs' verdicts compared: of 96 verdicts — every channel's pass or fail under all three references, the overall verdict, the identifiability classes and every pair's separation, and every family's detectability at its lowest and highest quantified severity at every sensor set — zero differ, and the largest move in any acceptance ratio is 0 at the six significant digits both runs publish, which is the precision at which the comparison can be made and not a claim that the two arithmetics agree exactly.

B 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; the extract is read only through that manifest, and the results file records the commit at which it was read. Four problems are left open by this report and are stated here so that they are not mistaken for settled questions. The day-to-day component of the refrigerant-side scatter is estimated from eight day blocks and is the least well determined quantity in the analysis. The condenser approach implied by the normal runs at full load disagrees with the machine's design value, which the design-approach form of the power law inherits. The mechanisms whose magnitude coverage falls under the alternative-mechanism program are representable and are not yet represented. And the smallest quantified severities of three fault families are not separable on this instrumentation, which no amount of computation will change.

C The protocol record

The protocol file was committed on 2026-09-22; its content hashes over line-feed-normalized bytes to

```
60e93d0b66953fe56a64fd49a7368edb
f42811f244622622608dc5250c3e9bba
```

and Table 19 carries the commits that fix it, the software it named, the dated correction to that software and the commit at which the computation ran. The computation refuses to start unless the protocol is committed and the software unmodified, and it records both. Table 20 enumerates everything that had been computed on normal operation beforehand.

Three facts about fault runs were known to the author before the protocol was fixed, from the benchmark's own published description, and are disclosed here because they shaped two prior widths and nothing else: at the deepest excess-oil severity the oil-sump temperature is recorded as rising by about 11.7 F, while the evaporator approach moves by about −0.05 F and the condenser approach by about +0.11 F. The first set the width, not the center, of the excess-oil magnitude prior; the second supports the point mass at no oil-to-evaporator coupling. No diagnosis, accuracy, calibration or detection outcome has been computed for any fault run.

Table 21 is a prediction, not a result: for each case of the benchmark's fault catalog it gives how often the model answers *not normal* when shown plants simulated at that case's recorded severity.

It is published here so that the companion report's measured outcomes can be scored against a column that existed before them. Cases whose outcome follows from the instrumentation alone — a condenser-flow fault read directly by the condenser flow meter, an oil fault read directly by the oil sump — are excluded and named.

Table 19: The record that fixes what was decided before the computation ran. The protocol file was committed on the date shown; the software it names was committed before it; one dated, purely additive correction to the software was committed after the protocol and before any value of the computation existed; and the computation itself ran at the last commit shown, with no change to the model, the protocol or the measurements in between.

What	Date	Commit or object
Protocol registered	2026-09-22	`b3e770b127659e3dbb9875f13da4403b1a3a4357`
Protocol content, as a stored object	2026-09-22	`0962ce9b679c43f6b88374d2d1deeb616020af1f`
Software the protocol named	2026-09-22	`6268dd93cea70a91559e0e9857d11917d573d922`
Software correction, dated	2026-09-22	`0aeb14c4e7571737a37aa0a29fff8c0f0feec060`
Computation on normal operation	—	`b3e770b127659e3dbb9875f13da4403b1a3a4357`

D Constants and reconciliation

Tables 22 and 23 list every constant the computation read and whether the protocol fixed it. Table 24 compares the validation quantities the protocol states to two decimals with the same quantities recomputed here at full precision.

Table 20: Everything computed on normal operation before the protocol was fixed; the list is part of the protocol file itself and is reproduced here in its own order, each entry's first clause only (an ellipsis marks a clause cut short). Nothing on it is a fault-run quantity, and the acceptance statistic this report tests — the day-block ratio on all three channels, under both forms of the test — is not on it.

# [index]	Quantity known before the protocol was fixed
1	the frozen R-134a saturation map and its residual
2	the conductance-law feature sets, chosen by day-block cross-validation on the benchmark normals
3	the point-level acceptance ratios TEA 1.377, TCA 1.244, TRC_sub 1.478 on the 18 benchmark normals in 9 day blocks, and a plain linear regression of TEA reaching 1.40
4	the point-level channel scatter 0.2069, 0.1842, 0.1802, 0.2800 K
5	the duplicate-sensor screen's two rejections
6	the gate's two known numerators
7	the evaporator resistance split that failed: a water-film share law for UA_e cross-validates at 1.910 F on TEA
8	the pre-service drifts of the run-mean residuals
9	every candidate law the model construction cross-validated leave-one-day-block-out on the 17 usable normals
10	the day-block CV of the laws as built into the implementation
11	the oil-step variants
12	the fixed-conductance power scans
13	the latent-approach UA fit
14	the identified structural quantities of the model construction: dkW/dT_c = 1.536 kW/K
15	the non-condensable offset of nc1 under the Dalton rule, about 0.28 K at the reference condition
16	the run-mean aliasing sets of the construction's test on the 18 benchmark normals at a condenser pool fraction of 0.20 and oil offsets of 1 and 3 K
17	the author's recomputation of those sets on the 17 usable post-break normals at the then-registered inputs
18	the metered flows of the post-break normals: normal1's FWE median of 215.08 gpm
19	the construction's runs-check script: every held-out normal of the 8 normal day-block folds diagnosed at L0 under each offset center, printing only the counts of runs that ran and were ...
20	a construction step's one RP-1043 access: the reference fold's exchanger
21	the seven composed nc prior numbers, computed by the author before protocol by the stages' own prior path on the reference fold's fitted exchanger and its screened normal ...

Table 21: A prediction, fixed before the companion report ran, of how often the model answers *not normal* for each case of the benchmark's fault catalog at the laboratory instrumentation. Each entry is computed from 60 draws of the model's own simulated plant at the severity the benchmark records for that case, on a normal template, with an exact binomial interval: no measurement of any fault run enters it. Twelve cases are excluded, either because the answer follows from the instrumentation alone — a water-flow fault read directly by that loop's flow meter, an oil fault read directly by the oil sump — or because the benchmark records no severity amount for the case. The companion report scores its realized answers against this column.

Case	Family	Severity [%]	Predicted probability [–]	Interval [–]
cf12	cf	12	1	[0.940371, 1]
cf20	cf	20	1	[0.940371, 1]
cf30	cf	30	1	[0.940371, 1]
cf45	cf	45	1	[0.940371, 1]
cf6	cf	6	0.666667	[0.533127, 0.783131]
eo14	eo	14	1	[0.940371, 1]
eo32	eo	32	1	[0.940371, 1]
eo50	eo	—	excluded	—
fwc10	fwc	—	excluded	—
fwc20	fwc	—	excluded	—
fwc20 alt	fwc	—	excluded	—
fwc30	fwc	—	excluded	—
fwc40	fwc	—	excluded	—
fwe10	fwe	—	excluded	—
fwe20	fwe	—	excluded	—
fwe30	fwe	—	excluded	—
fwe40	fwe	—	excluded	—
nc trace	nc	—	excluded	—
nc trace2	nc	—	excluded	—
nc1	nc	1	0.766667	[0.639617, 0.866163]
nc2	nc	1.8	0.966667	[0.884719, 0.995937]
nc3	nc	2.4	1	[0.940371, 1]
nc5	nc	5.6	1	[0.940371, 1]
rl10	rl	10	1	[0.940371, 1]
rl20	rl	20	1	[0.940371, 1]
rl30	rl	30	1	[0.940371, 1]
rl40	rl	40	1	[0.940371, 1]
ro10	ro	10	1	[0.940371, 1]
ro20	ro	20	1	[0.940371, 1]
ro30	ro	30	1	[0.940371, 1]
ro40	ro	40	1	[0.940371, 1]

Table 22: The first half, by name, of the 77 constants the computation read, with the value each took and whether the protocol file fixed it. A constant the protocol fixes is cross-checked against the model's own value before the computation starts, which stops if the two differ. Values shown as bounds are at the round-off floor; list-valued constants are truncated.

Constant	Value	Fixed in the protocol
charge_overfeed_share	$< 10^{-9}$	yes
charge_relative_starvation_slope	15	yes
charge_seal_fraction	0.15	yes
charge_smoothing_fraction	0.02	yes
condenser_pool_charge_fraction	0.3	yes
condenser_pool_charge_fraction_range	0.2, 0.44	yes
condenser_water_film_share	0.65	yes
condenser_water_film_share_range	0.4, 0.9	yes
condensing_pressure_drift_k	0.1	yes
condition_discrepancy	True	yes
condition_discrepancy_channels	TEA, TCA, TRC_sub, ...	yes
conductance_drift_covariance_root_tolerance	$< 10^{-9}$	yes
conductance_drift_ln_ua	0.02	yes
conductance_drift_rank_tolerance	1e-08	yes
conductance_drift_state_components	evaporator_fouling, condenser_fouling	yes
conductance_drift_step_ln_ua	0.0001	yes
design_approach_condenser_k	1.44444	yes
design_approach_evaporator_k	1.11111	yes
design_approach_log_sd	0.5	yes
design_conductance_load_exponent	0.35	yes
design_conductance_load_exponent_range	$< 10^{-9}$, 0.6	yes
detection_contradiction_level	0.05	yes
directional_deadband_multiple	3	yes
duplicate_sensor_tolerance_f	1	yes
eo_oil_offset_at_lowest_severity_k	1.232	yes
eo_quadrature_nodes	$< 10^{-9}$, 0.00555455, 0.0264008, ...	yes
eo_quadrature_parameter	oil_evaporator_loss_per_k	yes
eo_quadrature_point_mass	0.5	yes
eo_quadrature_representative	0.02	yes
eo_quadrature_support	$< 10^{-9}$, 0.08	yes
eo_quadrature_weights	0.5, 0.0869637, 0.163036, ...	yes
evaporator_water_film_share	0.6	yes
evaporator_water_film_share_range	0.4, 0.85	yes
excess_oil_evaporator_loss_per_k	0.08	yes
excess_oil_evaporator_loss_range	$< 10^{-9}$, 0.08	yes
excess_oil_evaporator_loss_zero_mass	0.5	yes
joint_offset_channels	TEA, TCA, TRC_sub, ...	yes
magnitude_derivative_step	1e-05	yes
map_gradient_tolerance	0.0001	yes

Table 23: The second half, by name, of the 77 constants the computation read, with the value each took and whether the protocol file fixed it. A constant the protocol fixes is cross-checked against the model's own value before the computation starts, which stops if the two differ. Values shown as bounds are at the round-off floor; list-valued constants are truncated.

Constant	Value	Fixed in the protocol
mode_search_starts	3	yes
nc_quadrature_nodes	0.0469101, 0.230765, 0.5, …	yes
nc_quadrature_parameter	noncondensable_segregated_	yes
nc_quadrature_point_mass	$< 10^{-9}$	yes
nc_quadrature_representative	0.5	yes
nc_quadrature_support	$< 10^{-9}$, 1	yes
nc_quadrature_weights	0.118463, 0.239314, 0.284444, …	yes
nitrogen_pressure_per_volume_percent_pa	6088.02	yes
nominal_evaporator_dry_fraction	0.1	yes
nominal_flooded_area_fraction	0.15	yes
noncondensable_diffusion_range	$< 10^{-9}$, 0.03	yes
noncondensable_segregated_share	0.5	yes
noncondensable_segregated_share_range	$< 10^{-9}$, 1	yes
noncondensable_system_to_bundle_volume_ratio	1.8	yes
noncondensable_system_to_bundle_volume_ratio_range	1.4, 2.5	yes
oil_sump_sensitivity_k_per_lb	0.4	yes
oil_sump_sensitivity_range_k_per_lb	0.27, 0.51	yes
outcomes_known_in_advance	3 structured entries	yes
overfeed_sensitivity_span	$< 10^{-9}$, 0.2	yes
pool_clearance_range	$< 10^{-9}$, 0.2	yes
quadrature_grid_span_prior_sd	-12, 6	yes
quadrature_points	200	yes
quadrature_steps_per_prior_sd	4	yes
reliability_bin_edges	0.125, 0.5, 0.8, …	yes
roundoff_bound	1e-09	yes
rule_sign_share	0.25	yes
spread_seeds	1043, 1044, 1045, …	yes
structural_fallback_share	0.25	yes
structural_hermite_points	5	yes
structural_interior_sd	4	yes
structural_quadrature_points	64	yes
structural_recentring	2	yes
structural_recentring_tolerance	0.25	yes
structural_truncated_points	12	yes
structural_window_sd	6	yes
superheat_authority_range	$< 10^{-9}$, 0.5	yes
tie_relative_tolerance	1e-06	yes
visibility_tolerance	1e-06	yes

Table 24: The validation quantities the protocol fixes, recomputed here at full precision under the same seed and the same draw counts and compared with the values the protocol states. A value agrees when it differs from the stated one by at most half a unit in the stated value's last decimal, and a count agrees when it is equal. Of 151 values compared, 151 agree; any that did not would be listed individually beneath its group, and none is.

Group of quantities	Compared [count]	Agreeing [count]
the quantities cited in the weakness classification	43	43
false alarms under a day-block conductance drift	3	3
identification, by fault family and sensor set	35	35
the self-recovery headline quantities	35	35
the calibration quantities	35	35

E Notation

The symbols used in more than one section are collected in Table 25.

Table 25: Symbols used in more than one section.

Symbol	Meaning
z	the state: every unknown of one fitted graph
$r_k(z)$, w_k	the residual of relation k and its weight, an inverse variance
Λ, σ_i	the information matrix and an unknown's posterior standard deviation
Z_h	the evidence of hypothesis h; differences of its logarithm are in nats
UA_e, UA_c	the conductances of the evaporator and of the condenser
Q, P	the cooling load and the compressor power
T_e, T_s	the evaporating temperature and the condenser shell's saturation temperature
η, η_0, e_1, e_2, e_3, P_0, x	the compressor efficiency, its coefficients, the constant power and the log load
T_{ewo}, T_{cwi}	the leaving chilled-water and entering condenser-water temperatures
M, u	a fault's magnitude and its natural logarithm
ξ	a structural constant integrated out by quadrature: the gas's segregated share or the oil coupling
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
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
σ_B, σ_R, σ_P	the day-block, run and point components of the scatter
σ_W	the within-block standard deviation of a channel
CV_c, E_c	a channel's cross-validated error and the error expected of a correct model
R_c, κ, b_c	the acceptance ratio, its limit, and the misfit bound
$\#\{\cdot\}$	the number of draws satisfying the condition in the braces
$\hat{h}_j$, h_j	the answered and the simulated hypothesis of draw j
D, $\mathcal{D}_f$, D_f	the number of simulated draws; the draws of family f and their number
$\hat{f}_s$, ρ_f, γ	the false-alarm rate, the identification rate and the coverage
$\bar{p}$, a	the mean stated confidence and the accuracy
s_f, ϑ_{fg}	a fault's whitened signature and the angle between two of them
$I(M)$, $M_{\min}$	the information the readings carry about a magnitude, and the smallest detectable one

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.

[Gordon and Ng 1995] Gordon, J. M., Ng, K. C. (1995). Predictive and diagnostic aspects of a universal thermodynamic model for chillers. International Journal of Heat and Mass Transfer 38(5), 807–818. doi:10.1016/0017-9310(94)00208-D.

[Henderson 1953] Henderson, C. R. (1953). Estimation of variance and covariance components. Biometrics 9(2), 226–252. doi:10.2307/3001853.

[Talts et al. 2018] Talts, S., Betancourt, M., Simpson, D., Vehtari, A., Gelman, A. (2018). Validating Bayesian inference algorithms with simulation-based calibration. arXiv:1804.06788.

[Tierney and Kadane 1986] Tierney, L., Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. Journal of the American Statistical Association 81(393), 82–86. doi:10.1080/01621459.1986.10478240.

[Loeliger 2004] Loeliger, H.-A. (2004). An introduction to factor graphs. IEEE Signal Processing Magazine 21(1), 28–41. doi:10.1109/MSP.2004.1267047.

[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.

[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). Sensor-bias and flow inference in chilled-water plants from the water-side measurements of a laboratory chiller under a pre-registered protocol. Technical report.