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

Ricardo Marquez, Ph.D.
rmarquez123@gmail.com

September 2026

Abstract

Chilled-water plants are controlled and audited through a few water-side sensors whose biases go undetected, and a biased flow meter or thermometer pair misstates the plant's load and efficiency. This report asks what one chiller's water-side sensors can reveal about their own biases and about the evaporator flow, and it is an inference study: a conservation-factored chilled-water model with a measurement layer for the sensors is fitted by maximum-a-posteriori estimation, and its intervals come from the Laplace approximation, under a protocol whose thresholds, draw counts and pass rules were committed before any result. On the real data of ASHRAE RP-1043, with the evaporator flow meter withheld, the model recovers four measured flow reductions of 9.3 % to 36.0 % to within 0.79 % to 2.29 %, with simultaneous intervals of ±2.2 % to ±3.4 % that exclude the unfaulted flow in all eight cases. Seven of the eight registered intervals contain the measured flow; the eighth, fwe40 under the additive offset structure, ends 0.099 % (0.137 gpm) below it, so the registered all-or-nothing recovery rule fails. On synthetic data from a declared generator, the supply-to-return bias contrast is calibrated (its 90 % intervals cover at 0.9035) and detectable from 0.5 K; a flow-meter scale bias is not detectable at any level up to 5 %, where its power is 0.228; the common-mode temperature bias is not recoverable; and model adequacy cannot be posed without an independent load measurement. Every negative result was anticipated in the protocol before the run: the synthetic ones as expected outcomes, the recovery miss as a stated risk of the all-or-nothing rule. The synthetic arm used 19,500 draws.

1 Introduction

Central chilled-water plants remove the heat of large buildings, campuses and data centers, and they are operated, commissioned and audited almost entirely through a few water-side sensors: a supply and a return thermometer and a flow meter on the chilled-water loop, the same on the condenser loop, and a power meter on each chiller. The load a plant serves and the efficiency it achieves are computed from those readings, so a thermometer pair that drifts apart by a fraction of a kelvin, or a flow meter that reads a few percent high, misstates both, and the controls that act on them inherit the error. Katipamula and Brambley (2005) review the fault detection and diagnosis methods that building systems use, among them the detection of faulty sensors. The laboratory data set most used to study chiller faults is the one Comstock and Braun (1999) produced for ASHRAE research project RP-1043, in which a 90-ton centrifugal chiller was run through normal operation and seven faults at four severities each; it carries the water-side sensors of a real plant and, for the faults that reduce a water flow, a measured value of the changed quantity.

Three lines of work frame this report. Ng, Yik, Lee, Lee and Chan (2020) detect and diagnose biases of HVAC plant sensors, chilled-water temperatures among them, by Bayesian inference over bias hypotheses on a multi-chiller plant. Sun, Wang and Shan (2022) quantify the systematic and random errors of the flow meters of a central cooling system with several water-cooled chillers, placing energy-balance models as constraints on a Markov chain Monte Carlo sampler; this is the closest published statement of the construction used here, conservation entered as a factor of the inference rather than checked afterwards. Chong, Lam, Pozzi and Yang (2017) calibrate building energy models by Bayesian inference with a model-discrepancy term in the sense of Kennedy and O'Hagan (2001), and report parameter posteriors alongside it. On the modeling side, the thermodynamic chiller law of Gordon and Ng (1995) relates a chiller's power to its load and temperatures, and the factor-graph formulation (Loeliger, 2004) and the Laplace approximation (Tierney and Kadane, 1986) are the standard tools of the estimator used here. None of these constructions is new here. This report restates the bias-and-flow problem of the first two lines on a different data set, one laboratory chiller with a single meter per water loop, and it does not compare its numbers with theirs: their plants, sensors and noise differ from this one, and their reported thresholds are cited as context only.

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. Domains enter through plugins that supply their laws, units and file formats; the inference and the bookkeeping are shared. This study uses its HVAC plugin to write the chilled-water loop of the RP-1043 chiller, one energy balance per steady-state point, with a measurement layer that carries a supply-to-return bias contrast, a common-mode temperature bias and a flow-meter scale bias as named unknowns, and it fits that model to real and to synthetic readings. Section 3 states the framework and each of these constructs precisely.

The question of this report is what a single chiller's water-side sensors can and cannot reveal about their own biases and about the evaporator flow when the flow meter itself is withheld. It is an inference study. The construction fits the conservation-factored model to the real RP-1043 reduced-flow runs, with the flow meter and every channel computed from it withheld, and to synthetic draws from a declared generator in which biases of known size are planted. The design, thresholds and pass rules were fixed in a protocol file committed before any result was computed, in the manner that pre-registration prescribes (Nosek, Ebersole, DeHaven and Mellor, 2018), and every verdict below is read against that file, including the ones that fail. Section 2 describes the chiller, the data and the model, Section 3 the estimator, the error model, the synthetic design and the protocol, Section 4 the results, Section 5 their reasons and limits, and Section 6 the answers; the appendices hold the reproduction recipe, the conservation audit and the registration record.

2 System and data

The chiller. Figure 1 draws the system. The RP-1043 chiller is a 90-ton centrifugal machine with a flooded evaporator and condenser, installed in a laboratory whose facility supplies a controlled heat load to the chilled-water loop and a controlled heat sink to the condenser-water loop (Comstock and Braun, 1999). The chilled-water loop carries two thermometer pairs: the chiller's own, TEI and TEO at the evaporator inlet and outlet, and a duplicate facility pair. This report uses the

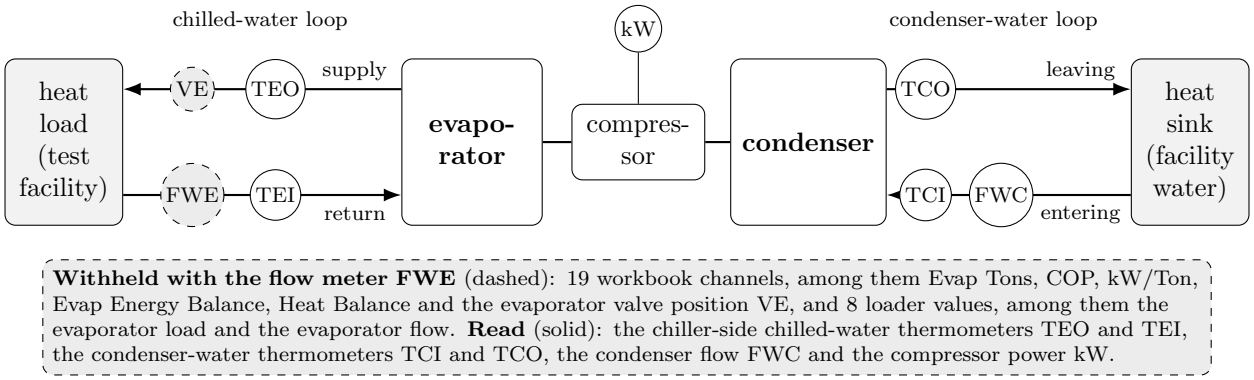


Figure 1: The chiller and the two water loops of this report, with the sensors it reads and the ones it withholds. Chilled water leaves the evaporator (supply, TEO), takes up the test facility's heat load and returns (TEI); the refrigerant cycle moves that heat and the compressor's power into the condenser, whose water loop carries it to the facility's heat sink (entering TCI, leaving TCO, flow FWC). The evaporator flow meter FWE is the quantity to be recovered, so it and every channel computed from it are withheld from the inference (dashed); the flow must then be reached through the condenser side.

chiller-side pair, because those are the sensors whose bias a plant operator would need to know; the facility pair enters only to set the thermometer noise, as half the variance of the pairwise disagreement, and never as a second observation. The flow meter FWE on the chilled-water loop is the quantity to be recovered. The condenser loop carries its thermometers TCI and TCO and its flow meter FWC, and the compressor its power meter.

The data. Every run used here holds 27 steady-state operating points, each a mean over 5 samples stored to 4 significant digits. The real arm uses the four reduced-evaporator-flow runs fwe10, fwe20, fwe30, fwe40 and the normal runs. Two further reduced-flow runs, `fwe20 alt--unsteady test5` and `fwe40 alt--unsteady test8`, are excluded because their workbook comments record that steady state was never reached. The facility's condenser-water flow was corrected between 25 Aug and 2 Sep 1999, and every normal run after the correction reports the same heat imbalance: the evaporator load plus the compressor power exceeds the heat the condenser rejects by 6.18 % of the latter on average over the 17 post-correction normal runs, with a standard deviation between runs of only 0.32 % against a within-run standard deviation of 2.3 % to 3.0 %. The one normal run before the correction, normal dpv, balances to 0.01 %. The imbalance is a property of the facility, not of any fault, and the error model of Section 3.5 is built to absorb it.

Table 1 lists which channels the inference reads. Withholding FWE withholds every workbook channel computed from it: the evaporator tonnage, the coefficient of performance, the energy balances built on them, and also the evaporator valve position VE, which tracks the flow closely. In all, 19 workbook channels and 8 values of the data loader are withheld, and a test of the study asserts, before any fit, that none of them reaches the likelihood or the data screen.

Screen and pairing. Before any fit, the readings of the scored and normal runs pass a screen on the channels of Table 1: a physical range and a within-run robust z of 6. Table 2 lists the seven points it rejects out of 594, all on the condenser flow, in two runs; the normal cf3 rejection at test 16 is the condenser-flow drop that the workbook comment for that run records, and the others lie in the one pre-correction normal run. FWE is never screened, because screening the truth would

Table 1: Channels of the real arm by use. The likelihood reads only the chiller-side chilled-water thermometers, the condenser side and the power; the flow meter FWE is read on normal runs only, to calibrate the error model, and on a scored run only as the truth. The withheld rows list every workbook channel and loader value computed from FWE or able to reveal it.

Use	Channels
Likelihood	`TEI`, `TEO`, `TCI`, `TCO`, `FWC`, `kW`, `compressor_running`
Data screen	`TEI`, `TEO`, `TCI`, `TCO`, `FWC`, `kW`
Calibration only (normal runs)	`FWE`
Withheld channels (19)	`Evap Tons`, `Shared Evap Tons`, `Building Tons`, `Evap Energy Balance`, `COP`, `kW/Ton`, `FWE`, `Tsh_suc`, `Heat Balance (kW)`, `Heat Balance%`, `Active Fault`, `PO_net`, `VSS`, `VSL`, `VH`, `VM`, `VC`, `VE`, `VW`
Withheld loader values (8)	`evaporator_load_w`, `cop`, `evaporator_water_flow_m3_per_s`, `as_chiller_test_point`, `as_chiller_test_point.load_w`, `benchmark_chiller_test_points`, `power_curve`, `power_curves`

Table 2: The 7 points the data screen rejected out of 594 screened, each beyond a within-run robust z of 6 on the channel shown (flows in gpm).

Run	Test	Channel	Reading	Robust z
normal cf3	16	`FWC`	255.3	8.13
normal dpv	4	`FWC`	205.08	9.36
normal dpv	7	`FWC`	205.54	9.18
normal dpv	8	`FWC`	205.8	9.08
normal dpv	13	`FWC`	211.86	6.71
normal dpv	16	`FWC`	205.24	9.30
normal dpv	25	`FWC`	205.74	9.11

select points by the answer. Each scored run is paired with a normal run by the registered date rule: the nearest earlier normal run in the same flow regime, never crossing the flow correction, with a same-date tie taking the earlier run in the workbook order. Table 3 lists the four pairs; fwe40 shares its date with normal2 and takes it by the tie rule, while the three other runs share normal1.

The model. Figure 2 shows the model as a factor graph. At each point the unknowns are the supply and return temperatures, the flow and the load; the energy balance ties them, the two thermometers read the temperatures through a measurement layer with a contrast and a common mode, and the load carries a prior formed from the condenser side. With the flow meter withheld, the condenser balance is the only route to the flow, and the common mode, which cancels from the temperature difference, reaches no factor at all.

The synthetic data. The synthetic arm draws readings from a declared generator that mirrors one run: 27 points per draw, evaporator temperature differences from 1.47 to 6.36 K, supply temperatures with a standard deviation of 2.28 K uncorrelated with the difference, a nominal flow of 13.5498 kg/s and a mean supply temperature of 280.504 K, and the condenser side of the normal run normal1 (268.56 gpm, 49.656 kW). Thermometer noise is 0.039 K on the supply, 0.117 K on the return, 0.371 K and 0.138 K on the condenser pair, each half the variance of the measured

Table 3: Pairing of the four scored runs with a normal run by the registered date rule. The deepest severity shares its test date with its normal run and takes it by the same-date tie rule.

Run	Date	Paired normal	Its date	Days	Rule applied
fwe10	1999-09-17	normal1	1999-09-16	1	nearest earlier normal run in the same flow regime
fwe20	1999-09-18	normal1	1999-09-16	2	nearest earlier normal run in the same flow regime
fwe30	1999-09-19	normal1	1999-09-16	3	nearest earlier normal run in the same flow regime
fwe40	1999-09-21	normal2	1999-09-21	0	normal run of the same date, earlier in the run index

duplicate-sensor disagreement; the flow noise is 0.2 % of reading, the within-run variation of FWE on the normal runs (0.17 to 0.23 percent); the power noise is one percent of reading, a stated assumption, since the data set records no power-meter class. The synthetic arm is labeled as such everywhere, and its numbers are never pooled with the real arm's.

3 Methods

3.1 The chilled-water loop and its measurement layer

At each steady-state point i the chilled water carries the load Q_i at the flow $\dot{m}_i$ between the supply temperature $T_{s,i}$ and the return temperature $T_{r,i}$. The thermometers read these temperatures through a measurement layer in which the contrast δ separates the two sensors' biases and the common mode μ moves both, and the flow meter reads the flow with a scale bias β_m,

$$Q_i = \dot{m}_i \, c_p \, (T_{r,i} - T_{s,i}), \tag{1}$$

$$y_{s,i} = T_{s,i} + \mu - \delta/2 + \varepsilon_{s,i}, \tag{2}$$

$$y_{r,i} = T_{r,i} + \mu + \delta/2 + \varepsilon_{r,i}, \tag{3}$$

$$y_{m,i} = \dot{m}_i \, (1 + \beta_m) + \varepsilon_{m,i}, \tag{4}$$

with $c_p = 4186.8$ J/(kg K) and a density of 998.554 kg/m^3, the workbook's own constant water convention, and each ε a zero-mean Gaussian reading error. The per-sensor biases follow as $b_s = \mu - \delta/2$ and $b_r = \mu + \delta/2$. Because the balance (1) sees only the difference of the two temperatures, μ cancels from every water-side factor, and δ enters exactly as an error of the temperature difference. The biases carry Gaussian priors of mean zero and standard deviations 0.5 K for δ and μ and 3 % for β_m.

3.2 The identification route

The load at each point is not measured. It is reached through the condenser: the condenser water rejects $\hat{Q}_{c,i}$, the product of its mass flow, c_p and its temperature rise from TCI to TCO, and that heat is the evaporator load plus the compressor power P_i. The load therefore carries the prior

$$Q_i \sim \mathcal{N}\big(\hat{Q}_{c,i} - P_i, \; s_Q^2\big), \tag{5}$$

the same construction on real and synthetic readings, so that the prior is never centered on the truth. Its width s_Q is derived rather than chosen, as the first-order propagation of the condenser

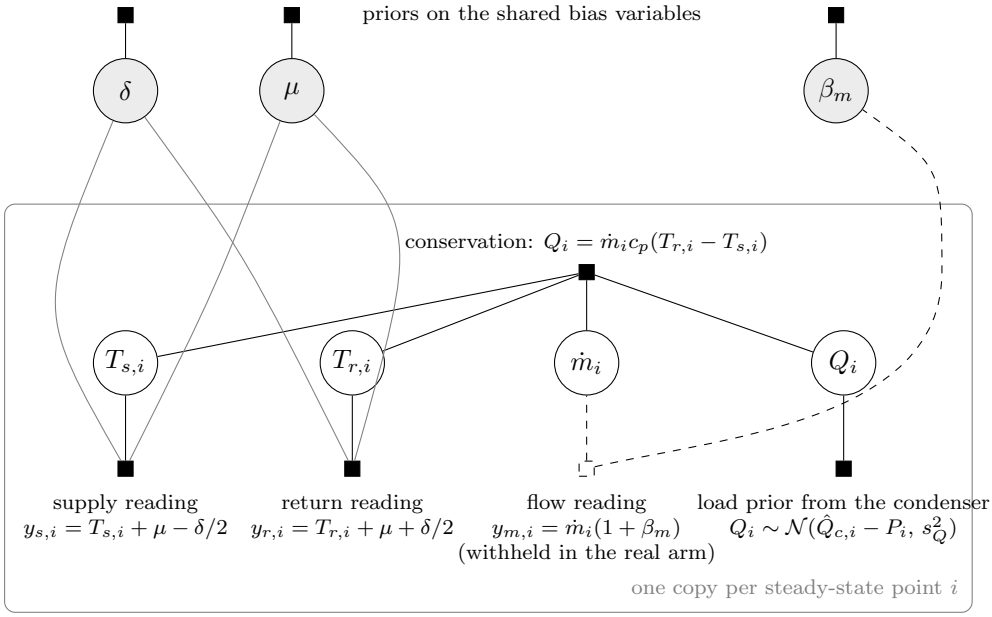


Figure 2: The model as a factor graph. Circles are unknowns and black squares are the relations (factors) that tie them; the shaded unknowns are shared by every point of a run, the others are repeated for each steady-state point i. At each point the chilled-water energy balance ties the load Q_i, the flow $\dot{m}_i$ and the supply and return temperatures; the two thermometers read those temperatures through the measurement layer, a contrast δ and a common mode μ; and the load carries a prior centered on what the condenser side implies, its heat rejection $\hat{Q}_{c,i}$ minus the compressor power P_i. The flow reading (dashed) carries the meter's scale bias β_m in the synthetic arm and is absent in the real arm, where the flows of a run are one unknown and the flow is reached only through the condenser-side load prior. The common mode μ cancels from $T_{r,i} - T_{s,i}$, which is why no water-side factor can see it.

thermometer and power noise at the condenser flow $\dot{m}_c$,

$$s_Q^2 = (\dot{m}_c\, c_p)^2\, (\sigma_{\mathrm{TCI}}^2 + \sigma_{\mathrm{TCO}}^2) + \sigma_P^2, \tag{6}$$

which gives $s_Q = 28.0439$ kW per point on the synthetic design. With FWE withheld, the evaporator flow is identified only through (1) and (5): the condenser side fixes the load, the chiller-side thermometers fix the temperature difference, and the balance returns the flow.

3.3 The chiller law, excluded

A chiller law such as that of Gordon and Ng (1995) relates the compressor power to the load and to the chilled- and condenser-water temperatures, and it is the only factor through which the common mode μ could be seen. It is deliberately left out of the likelihood. Fitted to the normal runs of this chiller and tested on a held-out run, its power error is 2.28 kW, which at the measured sensitivity of the power to the chilled-water temperature, 1.292 kW/K, is worth 1.77 K of chilled water, several times the biases of interest; its error is set by the operating condition rather than by noise, and its chilled-water slope differs from the data's by a factor of 1.7 to 3.1. Entered as a factor, it would attribute its own error to the sensors.

3.4 The probabilistic model

The relations above are written as residuals on a graph, and the engine fits them together; this subsection states the general model they belong to. This study compares no hypotheses by their evidence: its products are the maximum-a-posteriori state and the intervals built from the posterior standard deviations. 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,

$$p(z \mid y) \propto \prod_k \exp\!\big(-\tfrac{1}{2}\, w_k\, r_k(z)^2\big). \tag{7}$$

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:

$$C(z) = \tfrac{1}{2} \sum_k w_k\, r_k(z)^2, \tag{8}$$

$$z^* = \arg\min_z C(z), \tag{9}$$

$$z \leftarrow z - (J^\top W J)^{-1} J^\top W\, r(z) \quad \text{until } \|J^\top W\, r(z)\| < \epsilon. \tag{10}$$

Near the minimum the cost (8) 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:

$$p(z \mid y) \approx \mathcal{N}\big(z^*, \Lambda^{-1}\big), \tag{11}$$

$$\Lambda = J^\top W J, \tag{12}$$

$$\sigma_i = \sqrt{(\Lambda^{-1})_{ii}}. \tag{13}$$

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,

$$\log Z = -C(z^*) + \tfrac{d}{2} \log 2\pi - \tfrac{1}{2} \log\det \Lambda + \sum_k \tfrac{1}{2} \log \frac{w_k}{2\pi}. \tag{14}$$

Differences of $\log Z$ between hypotheses, from (14), 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.

3.5 The error model of the real arm

On the real data the balance does not close even on a normal run (Section 2), so each fit carries a correction of the facility's imbalance, in one of two registered structures that the normal runs cannot tell apart: a flow-scale ratio κ and an additive offset of the temperature difference, which is the contrast δ of the measurement layer,

$$Q_i = (1 + \kappa)\, \dot{m}\, c_p\, \Delta T_i, \tag{15}$$

$$Q_i = \dot{m}\, c_p\, (\Delta T_i - \delta), \tag{16}$$

with ΔT_i the observed chiller-side temperature difference and $\dot{m}$ the one flow of the run. Both are run on every scored run and reported side by side. Their error model is a run-level imbalance effect plus point noise, estimated on a fold of normal runs grouped by contiguous-day block: the post-correction normals minus the scored run's own day block. The correction's prior is centered on the paired normal run's effect with a standard deviation of $\sqrt{2}$ times the between-block standard deviation of block-mean effects, the spread of a difference of two run effects; the point noise is the fold's pooled within-run standard deviation of the same residual, and it is carried once, as the whole per-point noise of the balance, because the thermometer noise alone would imply about 7 kW of scatter per point against the 4.7 to 5.9 kW measured. For the reduced-flow runs the fold is all 17 post-correction normals in 8 day blocks (458 points); the paired normal's effect is -0.0808135 with a prior standard deviation of 0.0053199 and a point noise of 0.0321964 for the flow-scale ratio, and 0.268188 K, 0.0232385 K and 0.104038 K for the offset.

Each run's interval for the flow is simultaneous over the m runs and structures scored together, at the family level $\alpha = 0.05$, and uses a Student t quantile on one fewer degree of freedom than the fold has day blocks, because so few blocks estimate the between-block variance,

$$\dot{m} \in \hat{\dot{m}} \pm t_{\nu,\, 1-\alpha/(2m)}\ \sigma_{\dot{m}}, \tag{17}$$

with $\hat{\dot{m}}$ the maximum-a-posteriori flow and $\sigma_{\dot{m}}$ its posterior standard deviation (13). For the reduced-flow runs $m = 8$, the level is 0.99375, $\nu = 7$ and the quantile is 3.85523. No widening for the flow meter's own uncertainty is applied, because the data set records no meter class for FWE. A sensitivity analysis, reported beside every registered fit and entering no verdict, lets the run effect grow linearly with the temperature difference, because the within-run balance ratio correlates with it (correlation 0.34 to 0.56) in every post-correction normal run.

3.6 The synthetic generator and the draws

Each synthetic draw is one fit of the model of Figure 2 with the flow reading present, to readings generated from the design of Section 2 with every bias drawn from its prior (coverage), set to zero (the detection null), or set to a grid magnitude with the others zero (detection power). Coverage uses 2,000 draws per sensor set, the detection null 1,000, and every grid point 500 draws. The seeds derive from a registered root, 20260917, by named streams. A draw whose fit fails is counted, never dropped; a gate with more than one percent failed draws is reported as not evaluated.

3.7 Metrics

Coverage is the fraction of draws whose 90 % marginal interval contains the planted value; the gate is the registered band [0.87, 0.93], which a true coverage of 0.85 passes with probability below one percent, and which a conservative posterior fails too. Beside it stands the *contraction* $1 - \sigma^2_{\text{post}}/\sigma^2_{\text{prior}}$ and an identifiability class set by the ratio $\sigma_{\text{post}}/\sigma_{\text{prior}}$ (*well* identified below 0.25, *weakly* from 0.25

to 0.75, *not* above 0.75), because an unidentified parameter covers trivially. *Detection* of a bias uses the posterior probability that its magnitude exceeds a practical threshold (0.25 K for δ and μ, 1.5 % for β_m); a draw is flagged when any channel's probability exceeds its decision level, and the levels are calibrated on the null draws as the maximum statistic, holding the family-wise false-alarm rate at 0.05. *Power* is the fraction of flagged draws at a grid magnitude, with its 95 % binomial interval (Clopper and Pearson, 1934); the *minimum detectable bias* is the smallest grid level whose interval's lower end reaches 0.8. The *adequacy* statistic flags a draw whose fit is worse than the model allows at $\alpha = 0.05$, and its false-alarm rate on no-fault draws is gated to the band [0.03, 0.08]. The real arm's *recovery* passes if and only if the measured FWE lies inside the interval (17) for every run and both structures; its *informativeness* requires every interval of the three deeper severities to exclude the paired normal's flow as well, so that no interval passes by being wide. Verdicts are computed from the six-significant-digit values printed here.

3.8 The non-aliasing rule

The adequacy gate needs an out-of-set mechanism, a way the plant can depart from the model that no modeled bias reproduces. An additive heat gain on the loop does not qualify: with the flow nearly constant within a run it is an exact alias of δ. The protocol therefore tries candidates in a fixed order, a loop loss proportional to the supply temperature's deviation from its mean and then to its square, with a coefficient γ of 500 W/K, and admits the first whose information about γ survives the other unknowns: one minus the R^2 of γ's signature on every bias and load unknown must reach 0.1. If none does, the adequacy gate is reported as not evaluable.

3.9 Protocol and amendments

The protocol file was committed on 2026-09-17, before any posterior, coverage, power or recovery number was computed, and the run script refuses to start while the file is uncommitted or modified. The results record the file's hash, its git blob and the commit that introduced its content (Appendix C). The protocol file carries five dated additions, each made before the first result and each additive: it adds values beside the registered ones and changes or removes none. Table 4 lists them with the reasons the file itself states. The first names the channels, the chiller-side thermometer pair and the screened channels, which the protocol had fixed only by their noise and thresholds. The second fixes the adequacy level at 0.05, the maximum-statistic apportionment, the two-sided detection statistic and the condenser-implied load-prior mean. The third replaces the load-prior width of 4 kW by the derived width (6), because a 4 kW prior on a construction uncertain by about 28.0439 kW would make every synthetic posterior overconfident; it states in advance that the identifiability expectations fall by one class as a consequence, turns the sensor-removal ladder from a gate into a reported design property, and fixes the contrast's detection grid, the non-aliasing rule and the condenser constants. The fourth fixes the nominal flow and supply temperature by the same rule as the condenser constants, and discloses that they were set after mechanics runs at fixture levels had shown neither out-of-set candidate passing (0.058 and 0.021 retained against 0.1). The fifth fixes the real arm's error model as stated above, the Student quantile, the fold of a fault run, the truths and the pairing of the deepest severity. The protocol's order check verifies, from the repository history alone, that the protocol commit precedes the first results commit, that every version of the file before that commit is additive relative to the first version, checked at every intermediate version so that a value changed and restored would still fail, and that every added value sits under a block dated no earlier than the registration. The rule admitting several pre-result versions dates from 2026-09-19: as first written, the check admitted exactly one version before the

Table 4: The protocol file's dated additions, read from the file itself, each made before the first result. *Entries* counts the values the block adds; no registered value is changed or removed by any of them.

No.	Dated	Block	Entries	Stated reason
1	2026-09-17	`data_screen.channels`	11	The original registration fixed the screen's thresholds but named no channels
1	2026-09-17	`model.channels`	11	The original registration named sigmas per sensor but never named the channels themselves
2	2026-09-17	`amendment_2_unspecified_values`	17	The build of the measurement layer and of inference.calibration found values the original registration left unspecified
3	2026-09-17	`amendment_3_build_findings`	47	Building the synthetic generator, the channel-dependency table and the conservation audit found values the registration left unspecified and one internal inconsistency
4	2026-09-17	`amendment_4_design_levels`	8	The synthetic-arm runner needs the nominal evaporator flow and the mean supply temperature, which the registration never fixed
5	2026-09-17	`amendment_5_real_arm`	17	Building the real-arm runner found choices the registration leaves open, each of which moves the recovery intervals

first result and rejected this study's five additions, and it took its present form after it had done so. This study is the one on which the change was made, and it is disclosed for that reason.

4 Results

The registered recovery fails. Seven of the eight simultaneous intervals contain the measured flow and all eight exclude the paired normal's flow, but the interval of fwe40 under the additive ΔT offset structure ends at 137.963 gpm, 0.137 gpm (0.099 %) below the measured 138.1 gpm, and the rule requires every interval. The remaining registered verdicts are listed in Table 5: the normal-run null passes; on the synthetic arm the contrast's coverage and detection pass, the flow scale's detection fails, and the other gates are not evaluable or not evaluated for reasons the protocol fixed in advance.

4.1 Flow recovery on the real data

Table 6 lists each run's measured flow, the estimate and simultaneous interval under both structures, the interval's half-width and the two registered tests, and Figure 3 draws the same intervals as deviations from the measured flow. The four runs reduce the flow by 9.29 %, 18.26 %, 27.93 %, 35.97 % relative to their paired normals. Under the flow-scale ratio the estimates lie 0.79 % to 1.74 % from the measured flow, all above it; under the additive offset 1.06 % to 2.29 %, all below it, the deviation growing with severity. The half-widths are 2.25 % to 3.43 % of the estimate, the flow-scale intervals being the wider, and every interval lies far from the paired normal's flow, so the intervals are informative. The one miss is the offset structure on the deepest severity, where the estimate is lowest relative to the truth and the interval narrowest in absolute terms.

Table 5: Every registered verdict of the study. The four verdict words are fixed in advance: *passed* and *failed* for a gate that was evaluated, *not evaluated* for a gate whose draws failed in excess of the registered one percent, *not evaluable* for a gate the design cannot pose.

Arm	Gate	Verdict
real arm	recovery (registered)	fail
real arm	normal-run null	pass
synthetic arm	`coverage/full/delta`	passed
synthetic arm	`coverage/drop_kW/delta`	not evaluable
synthetic arm	`coverage/drop_kW/beta_m`	not evaluable
synthetic arm	`coverage/drop_condenser_pair/delta`	not evaluable
synthetic arm	`coverage/drop_condenser_pair/beta_m`	not evaluable
synthetic arm	`coverage/drop_flow_meter/delta`	not evaluated
synthetic arm	`detection/delta`	passed
synthetic arm	`detection/beta_m`	failed
synthetic arm	`adequacy/false_alarm`	not evaluable

Table 6: Registered recovery of the evaporator flow, in gpm, for the four runs and both correction structures. *Measured* is the withheld FWE (the truth), ± is the simultaneous interval's half-width as a percentage of the estimate, *Contains* says whether the interval contains the measured flow, and *Normal* is the flow of the paired normal run, which the interval must exclude. *Scale* is the flow-scale ratio and *offset* the additive offset of the temperature difference.

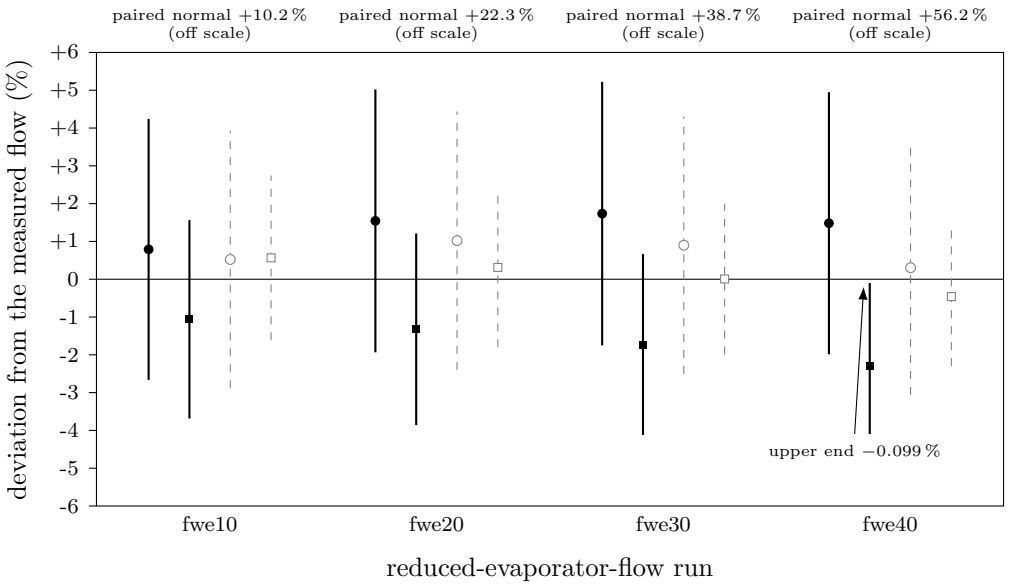
Run	Structure	Measured	Estimate	Interval	±	Contains	Normal	Excludes
fwe10	scale	195.2	196.741	[190, 203.482]	3.43 %	yes	215.2	yes
fwe10	offset	195.2	193.128	[188.005, 198.251]	2.65 %	yes	215.2	yes
fwe20	scale	175.9	178.616	[172.497, 184.735]	3.43 %	yes	215.2	yes
fwe20	offset	175.9	173.568	[169.109, 178.027]	2.57 %	yes	215.2	yes
fwe30	scale	155.1	157.793	[152.387, 163.199]	3.43 %	yes	215.2	yes
fwe30	offset	155.1	152.421	[148.711, 156.131]	2.43 %	yes	215.2	yes
fwe40	scale	138.1	140.147	[135.36, 144.934]	3.42 %	yes	215.684	yes
fwe40	offset	138.1	134.932	[131.901, 137.963]	2.25 %	no	215.684	yes

The null. The 17 post-correction normal runs pass through the identical pipeline, each against a baseline and a fold that exclude its own day block, scored separately with $m = 34$ (level 0.998529, $\nu = 6$). All 34 intervals contain the run's measured flow, with half-widths of 3.96 % to 5.09 %; the pre-correction run normal dpv has no partner in its regime and is left out. Table 11 in the appendix lists every run. The null passes, so the error model does not manufacture a flow change on runs that have none.

The temperature-difference sensitivity. Table 7 repeats the recovery with a run effect linear in the temperature difference. It is ungated. All eight of its intervals contain the measured flow and all exclude the paired normal; on fwe40 the offset estimate moves from 134.932 to 137.466 gpm and the interval to [134.93, 140.002] gpm. The sensitivity locates the miss; it does not change the registered verdict.

4.2 Calibration and detection on synthetic data

Table 8 lists coverage per sensor set and bias. With all sensors, the contrast's 90 % intervals cover in 1,807 of 2,000 draws, 0.9035 with a 95 % interval of [0.8897, 0.9161], inside the band, and the



black, solid: registered fits (circle: flow-scale ratio; square: additive ΔT offset); gray, dashed: ΔT-linear sensitivity, ungated
the horizontal line is the measured flow of each run

Figure 3: Recovered evaporator flow of the four reduced-flow runs, as a deviation from each run's measured flow. Each vertical bar is a simultaneous interval (the level of Section 3.7) and its mark is the estimate; the first two bars of a run are the registered fits, the last two the ungated sensitivity with a temperature-difference-dependent imbalance. An interval that crosses the horizontal line contains the measured flow. The paired normal run's flow, which each interval must exclude, lies far above the plotted range, by the percentage printed at the top. The one registered interval that stops short of the measured flow is marked with an arrow.

gate is passed. Its mean contraction is 0.923104, a posterior standard deviation 0.277 times the prior's, which places it among the weakly identified, and its mean interval width is 0.455934 K. The flow scale covers at 0.8960 but contracts by only 0.223681, is classified as not identified, and is therefore reported ungated. Without the return thermometer neither bias moves from its prior. Without the flow meter all 2,000 coverage fits fail as rank-deficient, because each point's flow is then unconstrained, and the gate is not evaluated; without the power meter or the condenser thermometers the registered load prior cannot be formed, since its mean and width both read the removed sensor, and those gates are not evaluable. Figure 4 draws detection power against the planted magnitude for the two gated channels. The decision levels, calibrated on 1,000 null draws, are 0.46722 for δ, 0.73652 for β_m and 0.617086 for μ, and they realize a family-wise false-alarm rate of 0.05 [0.0373354, 0.0653905]. The contrast is detected with power 0.988 [0.974, 0.996] at 0.5 K, its minimum detectable bias; at 0.35 K the power is 0.828 but the interval's lower end, 0.792, falls short of 0.8. The gate is passed. The flow scale reaches a power of only 0.228 [0.192, 0.267] at 5 %, the top of its grid; its minimum detectable bias lies above the grid and the gate is failed. The common mode, reported ungated, is flagged in at most 6.0 % of draws at any magnitude up to 3 K, which is the false-alarm rate: it is not recovered, as the protocol expected.

Table 7: Ungated sensitivity: the same recovery with a run effect linear in the evaporator temperature difference, flows in gpm, with the structures named as in Table 6. It is reported beside the registered fits and enters no verdict.

Run	Structure	Measured	Estimate	Interval	Contains	Excludes normal
fwe10	scale	195.2	196.219	[189.569, 202.869]	yes	yes
fwe10	offset	195.2	196.304	[192.049, 200.559]	yes	yes
fwe20	scale	175.9	177.695	[171.683, 183.707]	yes	yes
fwe20	offset	175.9	176.45	[172.741, 180.159]	yes	yes
fwe30	scale	155.1	156.495	[151.215, 161.775]	yes	yes
fwe30	offset	155.1	155.109	[152.014, 158.204]	yes	yes
fwe40	scale	138.1	138.521	[133.876, 143.166]	yes	yes
fwe40	offset	138.1	137.466	[134.93, 140.002]	yes	yes

Table 8: Coverage of the 90 % marginal intervals on the synthetic arm, per sensor set and bias, against the registered band [0.87, 0.93]; *Interval* is the 95 % binomial interval of the coverage. *Contraction* is the mean of $1 - \sigma^2_{\text{post}}/\sigma^2_{\text{prior}}$ over the draws and *Class* its identifiability class; a bias classified *not* identified is reported ungated, because an unidentified parameter covers trivially.

Sensor set	Bias	Draws	Coverage	Interval	Contraction	Class	Verdict
all sensors	δ	2,000	0.9035	[0.8897, 0.9161]	0.923104	weakly	passed
all sensors	β_m	2,000	0.8960	[0.8818, 0.9090]	0.223681	not	ungated
no power meter	δ	0	—	—	—	—	not evaluable
no power meter	β_m	0	—	—	—	—	not evaluable
no condenser thermometers	δ	0	—	—	—	—	not evaluable
no condenser thermometers	β_m	0	—	—	—	—	not evaluable
no return thermometer	δ	2,000	0.8815	[0.8665, 0.8953]	$\lvert c\rvert < 10^{-4}$	not	ungated
no return thermometer	β_m	2,000	0.9015	[0.8876, 0.9142]	$\lvert c\rvert < 10^{-4}$	not	ungated
no flow meter	δ	0	—	—	—	—	not evaluated
no flow meter	β_m	0	—	—	—	—	ungated

4.3 Adequacy

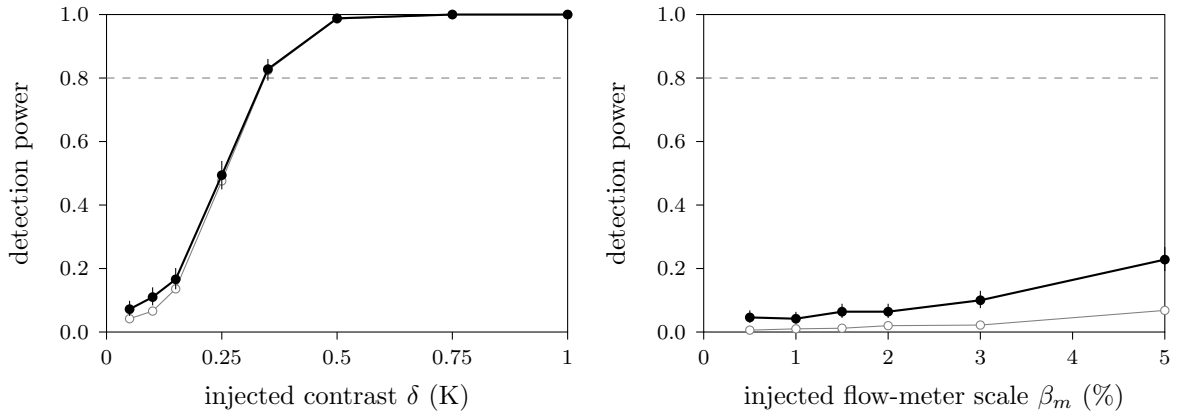
Neither out-of-set candidate passes the non-aliasing rule: the supply-linear loss retains 0.0588526 of its information and the supply-quadratic loss 0.0210109, both below 0.1, and in both the direction the data cannot see loads almost entirely on γ (0.981416 and 0.951934). The adequacy gate is therefore not evaluable. Its false-alarm rate on 1,000 no-fault draws is reported nonetheless: 54 flags, 0.054 [0.0408234, 0.069874], inside its band.

4.4 Sensor removal

Table 9 sets the identifiability class the protocol first expected for each sensor set beside the class expected once the load-prior width was derived and the class measured at the design point. With all sensors, δ is weakly identified ($\sigma/\sigma_{\text{prior}} = 0.277225$) and β_m not (0.882163), which is the amended expectation and one class below the original one for both. Without the return thermometer both stay at their priors, as both expectations said. The other three sets yield no measured class, for the reasons given above.

5 Discussion

The deepest severity. Two properties of the run explain the miss, and both were recorded before the run. First, the flow of fwe40 is not one number: FWE steps by 3.01 percent within



black: detection by any channel, the registered statistic, with its 95 % binomial interval; gray: the injected channel alone
dashed: the registered power of 0.8

Figure 4: Detection power of the two gated bias channels on the synthetic arm, each point from 500 draws. Left, the supply-to-return contrast δ; right, the flow meter's scale bias β_m in percent of reading. A level is detectable when the lower end of its interval reaches 0.8: the contrast first does so at 0.5 K, and the flow scale does not anywhere on its grid.

the run, and the registered truth is its mean, so a single run flow is fitted to points whose flow differs. Second, the imbalance grows with the temperature difference (correlation 0.34 to 0.56 on the normal runs), and the deepest severity runs at the largest differences, because a lower flow carries a comparable load across a larger temperature rise. An offset calibrated at the normal runs' temperature differences, where the additive structure (16) learns it, is then too small for this run, and the estimate falls 2.29 % below the truth. The ungated linear sensitivity removes the second cause and recovers the flow to within 0.46 %. The flow-scale structure (15), whose correction is proportional to the load and so grows with the temperature difference, covers all four runs. The miss is small, but the rule counts it, and it is reported as a failure.

Why adequacy cannot be posed. On this design the load at each point is an unknown with a prior about 28.0439 kW wide, so any heat the model does not describe — a gain on the loop, a loss that depends on the supply temperature — is absorbed as load. The non-aliasing rule measures exactly this: once the load unknowns are free, the supply-dependent losses keep a few percent of their information, and a statistic that asks whether the model fits cannot be expected to notice them. The question *no modeled fault explains this* needs an independent measurement of the load.

What another sensor would buy. The same load measurement is what the flow scale lacks. By the design arithmetic recorded with the protocol, the flow scale's posterior standard deviation would be 0.84 percent with the load known and 1.19 percent with a load prior of 4 kW, against 2.65 % at the derived width; a load-side heat meter, or a second flow meter on the chilled-water loop, would move β_m toward the practical threshold of 1.5 %. The common mode μ needs an absolute temperature reference: a third thermometer, or a chiller law validated to a small fraction of a kelvin.

Threats to validity. The real arm is one laboratory chiller, and each severity is one run, so the recovery results are single-run cases, not rates; three of the four runs share one normal run

Table 9: Identifiability of the two gated biases as sensors are removed: the class the protocol first expected (*Expected*), the class expected after the load-prior width was derived, one class lower (*Amended*), and the class measured at the design point, with the ratio of posterior to prior standard deviation. *Not constructible* marks a sensor set on which the registered load prior cannot be formed, because its mean and width both read the removed sensor; *no fit* marks a set whose design-point fit is rank-deficient.

Sensor set	Bias	Expected	Amended	Measured	$\sigma/\sigma_{\text{prior}}$
all sensors	δ	well	weakly	weakly	0.277225
all sensors	β_m	weakly	not	not	0.882163
no power meter	δ	well	weakly	not constructible	—
no power meter	β_m	weakly	not	not constructible	—
no condenser thermometers	δ	well	weakly	not constructible	—
no condenser thermometers	β_m	not	not	not constructible	—
no return thermometer	δ	not	not	not	1.00001
no return thermometer	β_m	not	not	not	1.00087
no flow meter	δ	well	weakly	no fit	—
no flow meter	β_m	not	not	no fit	—

as their baseline. The loop has a single flow meter, so the flow is identified through one balance, where a plant with several chillers has redundancy; the thresholds of the multi-chiller studies are not comparable. The protocol carries five additions, each made before the first result and each additive, but made by people who had by then built the machinery, and the fourth was made after mechanics runs at fixture levels; the file records this. The protocol was written before the model and its runners were built, and each addition fixes a choice that writing it did not foresee. The order check's rule for pre-result additions took its present form after it had rejected this study. The mechanism assumes an honest actor: the git checks prove order and identity, not that no result was seen.

6 Conclusion

A conservation-factored model of one chiller's chilled-water loop, fitted to the loop's water-side sensors with the flow meter withheld, recovers four measured evaporator-flow reductions of 9.3 % to 36.0 % to within 2.29 %, with simultaneous intervals that exclude the unfaulted flow in every case. Seven of the eight registered intervals contain the measured flow; the eighth misses it by 0.099 %, and the registered recovery fails. The miss sits at the deepest severity under the offset structure and is explained by an imbalance that grows with the temperature difference. On synthetic data the supply-to-return bias contrast is calibrated and detectable from 0.5 K, the flow-meter scale bias is not detectable up to 5 %, the common mode is not recoverable, and model adequacy cannot be posed without an independent load measurement. The next steps are a load-side heat meter or a second flow meter, which the design arithmetic says would make the flow scale identifiable, and a registered error model whose imbalance depends on the temperature difference.

A Reproduction

```
cd docs/terra_graph_engine/case_studies/foundation_hvac/plant_bias_and_adequacy
PYTHONPATH=../../../../../src python run.py
PYTHONPATH=../../../../../src python make_report.py
cd ../../../../..
```

```
python -m pytest -q tests/plugins/hvac/papers/test_plant_bias_and_adequacy_paper.py
```

`run.py` refuses to start unless the protocol file is committed and unmodified, runs both arms and writes `results.json` with the registration record embedded; `make_report.py` reads only `results.json` and the protocol file beside it and writes `report.tex`. The real arm reads the committed steady-state extract of RP-1043, whose manifest pins each file's hash. The test in the last line rebuilds the tex byte for byte and checks the registration record and the commit order.

B Conservation audit

The conservation relations of the model reproduce the data set's own balance columns on its raw readings. Table 10 lists the agreement over the 1,551 running points of the 59 runs, each point judged against the first-order worst case of the workbook's storage (4 significant digits per stored value and a 5-sample mean per steady state). Every column agrees at every point except the heat-balance percentage, which agrees at 1,549 of 1,551; the two exceptions are one point in each of `cf6` and `fwe40 alt--unsteady test8`, where the covariance of averaging a ratio over five samples, which the storage bound does not cover, exceeds the bound. By the workbook's own tolerance, the run-mean rule flags one run, `aborted eo86`, while the any-point rule flags 55 of the 59 runs and cannot separate them.

Table 10: The conservation audit: how many running points of each workbook balance column the conservation relations reproduce within the storage-derived tolerance.

Workbook column	Points reproduced	Running points
Cond Energy Balance	1,551	1,551
Cond Tons	1,551	1,551
Evap Energy Balance	1,551	1,551
Evap Tons	1,551	1,551
Heat Balance (kW)	1,551	1,551
Heat Balance%	1,549	1,551

C Registration record

The results were computed under the protocol file `docs/terra_graph_engine/case_studies/foundation_hvac/plant_bias_and_adequacy/preregistration.json`, registered on 2026-09-17. Its hash over LF-normalized bytes is

```
ee77e2dc4931108d6eaa2fac210f19ed
03f8087c33ae5a4ae39de94e24e3fe3c
```

its git blob is `686163a875df937dbf824fd9f96ee18e260cef26`, and the commit that introduced its current content is `1db087bb4ff3c20589127b9e6ec41a1ef4a8fa49`. The file's additions are dated 2026-09-17; Table 4 lists them. The test of Appendix A recomputes the hash and the blob from the committed file and runs the order check of Section 3 on the repository history.

Table 11: The normal-run null, flows in gpm: every post-correction normal run scored by the identical pipeline against a baseline and a fold that exclude its own day block.

Run	Baseline	Measured	Flow-scale interval	Contains	Offset interval	Contains
near normal1	normal1	215.314	[204.597, 226.559]	yes	[204.115, 221.255]	yes
near normal2	normal1	215.319	[205.008, 227.012]	yes	[203.941, 221.347]	yes
near normal3	normal1	215.714	[206.602, 228.778]	yes	[204.79, 221.746]	yes
normal b	normal r	214.903	[204.035, 225.647]	yes	[202.98, 222.49]	yes
normal cf	normal b	215.083	[204.74, 226.448]	yes	[203.893, 222.287]	yes
normal cf2	normal cf	214.809	[204.378, 225.406]	yes	[204.307, 222.269]	yes
normal cf3	normal cf	215.18	[204.012, 225.248]	yes	[203.581, 221.895]	yes
normal cf4	normal cf	214.938	[203.839, 224.813]	yes	[203.585, 221.489]	yes
normal cf5	normal cf	214.868	[204.326, 225.348]	yes	[203.458, 221.56]	yes
normal cf6	normal cf	214.922	[205.124, 226.228]	yes	[204.638, 223.45]	yes
normal eo	normal r	215.19	[204.679, 226.369]	yes	[204.983, 223.973]	yes
normal nc	normal cf6	214.49	[203.938, 223.55]	yes	[203.843, 220.915]	yes
normal r	normal2	215.607	[204.25, 224.53]	yes	[204.207, 221.039]	yes
normal r1	normal r	215.283	[205.49, 227.248]	yes	[205.217, 223.739]	yes
normal	normal1	215.212	[204.68, 226.65]	yes	[203.325, 220.991]	yes
normal1	normal	215.2	[203.495, 223.831]	yes	[203.2, 220.346]	yes
normal2	normal1	215.684	[204.991, 226.795]	yes	[204.241, 221.707]	yes

References

[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.

[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.

[Ng et al. 2020] Ng, K. H., Yik, F. W. H., Lee, P., Lee, K. K. Y., Chan, D. C. H. (2020). Bayesian method for HVAC plant sensor fault detection and diagnosis. Energy and Buildings 228, 110476. doi:10.1016/j.enbuild.2020.110476.

[Sun et al. 2022] Sun, S., Wang, S., Shan, K. (2022). Flow measurement uncertainty quantification for building central cooling systems with multiple water-cooled chillers using a Bayesian approach. Applied Thermal Engineering 202, 117857. doi:10.1016/j.applthermaleng.2021.117857.

[Chong et al. 2017] Chong, A., Lam, K. P., Pozzi, M., Yang, J. (2017). Bayesian calibration of building energy models with large datasets. Energy and Buildings 154, 343–355. doi:10.1016/j.enbuild.2017.08.069.

[Kennedy and O'Hagan 2001] Kennedy, M. C., O'Hagan, A. (2001). Bayesian calibration of computer models. Journal of the Royal Statistical Society Series B 63(3), 425–464. doi:10.1111/1467-9868.00294.

[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.

[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.

[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.