A measurement-and-verification-firewalled living lab in an occupied single-family home

Rolf Elmér, Independent Researcher
`rolf.elmer@aviverenavalis.se` • ORCID 0009-0007-6735-3256

Preprint — August 5, 2026

Supplementary material (FMEA + pre-registration record): doi:10.5281/zenodo.21420257,

Abstract

Field demonstrations of predictive control in residential buildings report *results* but rarely *how the multi-year evidence base was assembled*, at what cost, or how the running controller was kept from contaminating the very measurements used to judge it. We describe an occupied single-family home being built into a research instrument — a multi-year programme whose instrumented phase began in 2026, standing on a 4.5-year measurement record (2022–) — under three simultaneous, mutually hostile contracts: a frozen measurement-and-verification (M&V) baseline, a live model-predictive controller (MPC) under continuous development, and a research programme of pre-registered experiments. The contribution is architectural: a three-plane firewall (Control / Analytics / Research) with explicit boundary rules, a documented failure-mode-and-effects analysis (FMEA) of the cross-plane corruption paths, and a set of operating patterns — a byte-hashed frozen baseline, a shadow-first / deploy-dark / kill-switch deployment lifecycle, a living incident register, an experiment-window masking invariant, and pre-registration of analyses before data collection — each of which closes a specific corruption path. We give logged evidence that the pattern bears load in practice: a code-review fix plan closed under live operation, a dead-on-arrival actuation bug caught in shadow and never in production, a specification assumption falsified by a natural experiment and the guard shipped the same day, a genuine cross-plane contamination — a control input locked in a self-confirming echo loop — exposed by an independent source and closed the same day, and a bit-identical model replication logged as a fact. The design is $N = 1$ by intent: it is a method paper about how to build one lived-in home into a credible instrument, not a performance claim about one house.

1 Introduction

The case for predictive control of building energy is largely a simulation case: the canonical survey of model predictive control for buildings reports substantial simulated energy and emissions savings yet finds practical deployment still early-stage, and explicitly calls for a standardized performance-assessment methodology [10]. A recent review of field demonstrations of MPC and reinforcement learning for HVAC finds that most reported deployments use experiment protocols that may yield unreliable performance estimates, and that the literature focuses almost exclusively on energy cost while saying little about deployment process or cost [1]. Residential field demonstrations are especially thin: a 2024 survey could enumerate only twelve to the authors' knowledge, with reported savings spanning 2–49 % under widely varying evaluation practice [2]. The gap is not primarily one of algorithms. It is that a field study in an occupied home must satisfy several obligations at once, and the literature reports the outcome of that balancing act without documenting the act itself.

Three contracts that must coexist. An occupied home run as a research instrument is simultaneously bound by three contracts that pull against each other:

1. **A frozen M&V baseline.** To attribute savings credibly (IPMVP Option C [13,14], whole-facility weather-normalised), the baseline model must be *frozen* — fixed at a known date and never silently altered — or every savings number computed against it is invalidated.
2. **A live controller under continuous development.** The MPC that actually runs the house is edited constantly: new inputs, new horizons, bug fixes. Development velocity is in direct tension with the stillness the baseline demands.
3. **A research programme of pre-registered experiments.** Thermal-identification floats, price-forecast shadows, and demand studies run *on the same house*, each perturbing it in ways that, if pooled into the baseline or fed into the controller, would corrupt both.

Each contract needs the others' data, yet each can corrupt the others if they share state: a control tweak that alters the frozen baseline silently invalidates every published savings number; a research signal fed into the controller makes control behaviour depend on unvalidated science; experiment-contaminated data pooled into the attribution baseline biases the headline result — a building-energy instance of the data-leakage failure mode that drives irreproducibility across machine-learning-based science [17], here closed by the firewall [3].

Research question. *How does one instrument and control an occupied single-family home so that a frozen M&V baseline, a continuously-developed live MPC, and a pre-registered research programme can coexist over a multi-year horizon without corrupting one another?* This paper answers with an architecture and logged evidence from the programme's instrumented first months (2026), on a 4.5-year retrospective data foundation. We do not claim the house's numbers are representative — N = 1 is a deliberate feature of a method paper (§4). We claim that the *method* of keeping the three contracts separate is reproducible, and we show it surviving contact with real incidents.

2 Architecture

The system's core is a **three-plane firewall** (Fig. 1), formalised as an FMEA [3]. Each plane reads what it needs but **writes only its own state**; that single asymmetry is what keeps the three contracts from contaminating one another.

The three planes.

- **Control** — the MPC optimiser and device dispatch. It reads prices, forecasts, and learned model coefficients as *read-only inputs*, and writes *device commands only.*
- **Analytics (M&V / IPMVP Option C)** — the frozen whole-facility baseline (trained on the calendar-year 2025 baseline period, 2025-01-01 to 2025-12-31, n = 365 days; frozen 2026-05-15 at R^2 = 0.91, CV(RMSE) = 11.6 % in-sample) and weather-corrected savings attribution. It reads the frozen baseline and metered energy, and writes *the savings log only.*

- **Research** — thermal-ID, price-forecast shadow, demand and phase analysis. It reads *copies* of any data and writes *its own research logs only*.

The cross-contamination FMEA. The value of naming the planes is that the corruption paths between them become enumerable. Each mechanism below exists to close a specific path [3]:

- **Frozen, byte-hashed baseline** (closes Control → Analytics): the baseline artefact is byte-hashed at runtime and the server refuses to start on a byte change, with a pre-commit exclusion and an "M&V-firewall-safe" review gate on every deploy. A control change therefore *cannot* silently edit the baseline; re-freezing is a deliberate ceremony, never a side effect. Freezing a trained baseline and evaluating it in a later prediction period is the recognized M&V-2.0 practice for baseline integrity [15,16].
- **Read-only model coefficients** (closes the Analytics↔Control coupling): recursively-estimated (RLS) coefficients are read-only inputs to Control; the frozen baseline never re-reads them, so a re-fit cannot leak into attribution.
- **Research-only sensing** (closes Research → Control): no research-only signal enters the control objective; new inputs enter in *shadow-emit* (observation-only, kill-switched) before any control cutover, so control behaviour never depends on unvalidated science.
- **Experiment-window masking — "Invariant #8"** (closes Research → Analytics): every analysis checks its time window against the registry of active experiment windows and excludes contaminated data. Enforcement is **procedural, not mechanical**: a shared loader (`mv_invariant8`) and a written analyst procedure implement the rule, but no production write path applies it and analyses are not prevented from bypassing the loader. We state this precisely because our own audit showed the distinction matters — see §4 into a normal-operation baseline. Pre-registering analyses before data collection — endorsed for energy research by the field's own transparency-and-reproducibility guidance [18] and the broader pre-registration literature [21] — constrains researcher degrees of freedom on an inherently observational single-home programme.
- **Frozen cost model, observer-only tariff** (closes tariff drift → Analytics): the billing model is frozen and the vendor tariff API is observer-only with a drift warning; the invoice is authoritative, so a changed upstream tariff cannot re-base savings.
- **Vendor-response validation** (closes malformed upstream → any plane): vendor API loops validate response shape before processing, so a corrupt price/weather/device payload cannot silently drive control or analysis.

The shadow-first deployment lifecycle (Fig. 3). New control capability is never switched on directly. It runs first in **shadow** (computing and logging its intended action while the incumbent controller actuates), then **deploy-dark** (present in production but gated off behind a flag), then **kill-switched live** (actuating, with a single flag that reverts to the safe incumbent). The pattern converts a class of deploy-time failures into observations: a new actuator that would fail on its first live command instead fails visibly in shadow, where nothing depends on it (§3, C6).

The living incident register. Every incident is filed as a numbered risk-analysis note (an "NRA") carrying an explicit "firewall-safe" line — the one-line test *which plane does this change write to, and does it read only, never write, any other plane's state?* The register is a living FMEA, not a post-mortem archive: fixes and their verification are appended to the same note, and recurring defect families are cross-referenced (§3).

Pre-registration. Research experiments freeze their hypotheses and decision rules *before* data collection, with the git-commit timestamp of the plan as the scientific record of when it was frozen; no analysis decision may change after collection begins; the deposited plan enumerates eight hypotheses (H1–H8) with their frozen decision thresholds [4]. This is the research plane's own anticontamination guard — it prevents the experiment's analysis from drifting to fit the data the instrument produces.

3 Evidence that the pattern bears load

Architecture claims are cheap; the test is whether the pattern survives real operation. Each claim below is tied to a logged artefact in the project record, not to self-report.

A code-review fix plan closed under live operation. A gold-standard two-part review of the controller (≈17k lines) surfaced a prioritised defect list — a swallowed `NameError` that would have silently no-opped *all* relay actuation including the frost-guard fail-safe, a wrong-zone configuration leak re-introducing a floor-temperature over-clamp, a stale price-cache TTL — with the P1 correctness items fixed the same night and the rest triaged as tracked open work, all verified on the actual code path [5]. The firewall let this happen without a maintenance freeze: the fixes were Control-plane writes that provably did not touch the frozen baseline.

A dead-on-arrival actuation bug caught in shadow, never in production (the "C6" case). A new deferrable-load actuator was found to have never actuated at all — a POST-versus-GET defect meant every relay command had silently no-opped. Because the capability was gated behind the shadow-first lifecycle, the defect surfaced in shadow before any live dependence; the fix (r83.51) preceded the first real actuation, whose first genuine ON event is in the raw log with a timestamp, power, and price (445.9 W at 58.2 öre — the evidential point is the timestamp of first genuine actuation, not the magnitude). The bug was real, and the pattern is exactly why it was harmless.

A specification assumption falsified by a natural experiment — and metabolised the same day. A manual pool-topping event on 2026-07-14 drove a large draw through the same meter the DHW model assumed was DHW-only, falsifying a specification assumption (that garden/pool and hot water are separately metered). Rather than being absorbed as noise, the counter-evidence was turned into a guard the same day: a flow-signature classifier that discards non-DHW draws above a physical threshold, and a continuity guard shipped in the same release train (r83.58/59). The system metabolises disconfirming evidence into a standing check.

A metering premise falsified by its own validation record — and re-registered, not patched. A draw-subtraction term had been added to the tank estimator on the premise, written into the design specification, that the whole-house water meter sits on the tank's cold feed, so metered volume equals hot-water draw. The term was deployed **observation-only** behind the shadow-first lifecycle — three volume guards (an irrigation-magnitude discard, a multi-window continuity check, and a live-flow gate), no control writeback — and a before/after validation was pre-registered against a stated criterion: the day-mean estimator innovation (+4.2 K before the term) should fall toward the night level if the term is correct. Segmented on a week of parity data (686 usable records), the record answered a stronger question than it was asked. The day/night gap did **not** close; instead the whole innovation level swung from +7.03 K (pre-term day mean, $n = 290$) to −5.58 K and −12.75 K (post-term day and night, $n = 188/48$), re-zeroing at each restart and then drifting down again — −8.3 K after sixteen hours. Mechanism confirmed, magnitude wrong by roughly two to three times, in the systematic-over-subtraction direction. The falsifying evidence chain is the point: a midnight cold-appliance draw of +23 L dropped the estimator −2.9 K while an independent proxy moved −0.2 K (a cold draw applied as hot); whole-house daily totals ran 495 / 875 / 224 / 238 / 218 L against a realistic two-person hot-water demand of ≈ 80–140 L/day — an implied 50–70 % cold fraction, the term's first full day an 875 L day (three tank volumes, the heat pump not running continuously); and the three guards fired **zero** times over the whole live period, the sub-threshold multi-window pattern slipping all three, so the term subtracted 14.91 K of "draw" in 18.5 h (against $\approx$6.9 K plausible over the same window — the $\sim$2$\times$ over-subtraction shown, not asserted) while the estimate sank below its reheat floor. The stratification confounder is bounded honestly — a lump-average estimate versus a top-sensor proxy legitimately holds a $-3\ldots-5$ K offset during draws, physics not error, but it cannot account for -13 K or the midnight signature. Two disciplines made this a clean result: the observation-only deployment meant a two-to-three-fold error accumulating ≈ -6 K/day of phantom draw-down reached **zero** actuation, caught in shadow before the plant-model cutover gate it would have corrupted; and the pre-registration makes the negative verdict a recorded outcome against a frozen criterion, not a post-hoc reinterpretation. The remedy keeps the discipline — an interim hot-water fraction ($\approx$0.45) derived from **energy closure** rather than by eye, a pre-registered re-validation, and a draw-signature classifier as the proper fix — with the falsifying record labelled as its training material [5, on request].

A genuine cross-plane contamination, caught by an independent source (the echo loop). The firewall did not prevent this one a priori — which is exactly why it is reported. The client snapshot that feeds the controller its current solar-irradiance input was found to be sending back the *server's own previous output* rather than a fresh observation: the server echoed the value, used it in the control objective *and* in the online solar-gain learning, and the client wrote the echo back into the snapshot — a self-confirming fixed point. The value froze at 432 W/m^2 for roughly thirteen hours, including overnight, while an independently fetched satellite series ran live beside it; the divergence between the two is what exposed the loop. Root cause was traced to the echo topology (not to any single component), the loop was broken the same day by making the client send the fresh upstream observation, and a server-side source-of-truth hardening was registered as follow-up [6]. Two lessons entered the FMEA as an enumerated path: a read-back of one's own output is a contamination class that static plane-boundaries do not stop (it crosses no plane on paper — it *loops* within one), and the practical detector is redundancy — a second, independently sourced series against which a frozen value is visible. The failure ran for hours; the discipline's contribution was that it was *detectable*, *diagnosable from logs*, and *closed with a registered fix* rather than silently absorbed.

Bit-identical replication as a logged fact. Determinism is treated as a verifiable property, not an aspiration: model changes are staged so that the disabled path is *byte-identical* to the incumbent (verified by file comparison across dev/prod and behind default-off flags), so a refactor can be proven control-neutral before any behavioural change is enabled [5, 7]. When a claimed "deploy pending" line survived days past an actual deploy, a byte comparison of dev-vs-prod resolved it as stale paperwork rather than a live defect [7].

3b. The evidence the method produces (forward-reference to A4 — not owned here). From the same platform, the data engine produces publication-ready characterisations of the house — a weather-independent load baseline, an (absent) price-responsiveness and the resulting peak-shift headroom, and the calibration-window behaviour of the price-forecast model — the results that motivate a 48-hour demand MPC [forward-cite: A4, in preparation]. *These quantities are A4's core results and are referenced here only as evidence that the platform generates publication-ready output; they are not reported as results of this paper.*

4 Limitations

Procedural enforcement is not mechanical enforcement — and we learned the difference the hard way. In August 2026 an internal audit of Invariant #8 found that the masking loader had, for several weeks, been validating against *fabricated* experiment windows: its data root was a relative path, so it audited the development checkout and never the production tree, where four window directories existed of which only one was a real experiment. Separately, a code sweep showed that exactly one analysis script in the repository actually imports the shared loader; the remaining analyses honour the rule by convention. The over-masking removed 882 rows of otherwise valid normal-operation data from a 48-hour span, and — more insidiously — emitted *false* "this file stopped writing" warnings for a period when logging was in fact continuous.

No published result changed: the affected span carried no fitted parameter, and the experiment-week analyses read the experiment's own log rather than the masked production series. But the failure is worth reporting rather than repairing quietly, because it is the failure mode this architecture is supposed to prevent, and it illustrates the paper's own thesis: **a firewall that is documented but not executable is a firewall whose integrity depends on the discipline of whoever is at the keyboard.** A rule enforced by habit degrades silently and gives no signal when it has stopped being followed. Making Invariant #8 a hard gate in the production write path — so that masking is applied by code rather than remembered by an analyst — remains open work.

N = 1 — here a strength, not a weakness. This is a method paper about how to build *one* occupied home into a credible instrument; the architecture, not a population estimate, is the contribution. The house's own numbers do not generalise and are not claimed to; the *method* of firewalling three contracts does, and is the reusable object.

Private code until 2028. The controller source is not yet public. We state the opening plan explicitly: the evidence panels and tools underlying the reproducible-assembly method (§6) are already documented from public sources and committed tool docstrings, and a staged code release is planned for 2028; until then, claims are backed by logged artefacts and the committed reports cited here. We state the honest limit plainly, and do not exempt this paper from it: the incident records [5,6,7] are **author-attested** de-identified excerpts pending that release, so the independently-reproducible core today is the L2 evidence-panel method (§6, public/customer-owned sources and committed tooling) and the deposited FMEA and pre-registration [3,4] — the running controller implementation itself waits for 2028.

Operator dependence. The house has a single expert operator, and several guards (the shadow-to-live flip, the pre-registration ceremony, the incident filing) are operator actions, not autonomous ones. The architecture reduces but does not remove this dependence; scaling beyond N = 1 would require making these ceremonies enforceable without the operator in the loop.

5 Related work (L1)

We position against four points in the residential predictive-control and living-lab literature. The landscape is anchored by the canonical MPC-for-buildings survey [10], the reference occupied-building field demonstration [11], and the standard review of reinforcement learning for demand response [12]; the measurement side rests on IPMVP and ASHRAE Guideline 14 [13,14] and the automated-baseline accuracy line [15,16]. The MPC/RL field-demonstration review [1] documents that field demos overwhelmingly report end-state performance under a "near-exclusive focus on energy cost", and flags the missing deployment-process and cost half — the gap this paper's *process* documentation fills. Residential field demonstrations report final performance (e.g. an all-electric cold-climate house [2]; a panel/service-protection MPC field study [8]) but not the M&V-firewalled, incident-registered *process* by which the instrument was built and kept honest over years. Institutional test facilities (FLEXLAB-class [19]) provide controlled instrumentation but are not occupied homes under a frozen savings baseline; the living-laboratory literature [20] instruments occupied real environments, but for multi-domain comfort investigation rather than under an M&V-frozen savings baseline. Open field-infrastructure contributions such as the HEAPO heat-pump dataset [9] show the community's appetite for documented field data infrastructure — but supply *data*, not the operating discipline that keeps a live controller from contaminating the measurement.

The firewall also has an intellectual ancestry in two adjacent literatures. That feedback contaminates what can be learned from operational data is a classical result in system identification [23], and its consequences for building thermal models under closed-loop operation are well documented [24,25]; that literature, however, treats the bias as an *estimation* problem to be mitigated by experiment design, whereas we treat the symmetric *evaluation* problem — the controller biasing the measurement of its own savings — as an architectural one, solved by a standing firewall between the control, analytics, and research planes. Related concerns appear in demand response, where counterfactual baselines are known to be vulnerable to strategic manipulation by participants [26]; the three-plane firewall addresses the complementary threat in which the contaminating agent is not a strategic customer but the deployed controller itself.

Table 1: Positioning against the residential field / living-lab literature.

Work	Occupied home?	Frozen M&V baseline?	Pre-registered experiments?	Incident register?
MPC/RL field-demo review [1]	reviews mostly occupied demos	not analysed (flags protocol unreliability)	no (protocols often ad hoc)	no
All-electric cold-climate demo [2]	yes (single-family)	no (reports end-state savings)	no	no
Panel/service-protection MPC [8]	yes (single-family)	no (proof-of-concept field test)	no	no
Institutional living labs (FLEXLAB)	no (purpose-built)	partial (controlled, not IPMVP-frozen)	varies	facility logs, not methodological
HEAPO dataset [9]	yes (1,408 homes, data only)	n/a (dataset)	n/a	no
This work (A3)	**yes (occupied single-family)**	**yes (IPMVP Option C, frozen 2026-05-15, byte-hashed)**	**yes (git-timestamped plans)**	**yes (living NRA register + FMEA)**

Mini-search protocol: query "residential living lab occupied M&V" + the named anchors; each arXiv id and DOI verified against arXiv and CrossRef (see references). The table's discriminating columns are the four contracts this paper argues must coexist; to our knowledge no prior residential field study reports all four.

6 Reproducible evidence-panel assembly (L2)

A corollary contribution answers the review's deployment-cost gap [1] directly: the multi-year evidence base — feeding the demand-flexibility question where nameplate capacity overstates the controllable fraction [22] — is assembled at near-zero cost, reproducibly, from public or customer-owned sources. A three-to-four-year hour-resolved evidence base for the occupied house is built from three sources, all joinable on the same `(date_local, hour_local)` key:

1. **Consumption** — the grid operator's "Mina sidor" hourly export (free; the customer's legal right to their own metered data). The parser handles the Swedish export format end-to-end (UTF-8 BOM, semicolon delimiter, decimal comma, `(CET|CEST)` tags, DST gaps and duplicate hours). Tool: `merge_consumption_hourly.py`.
2. **Weather** — the Open-Meteo ERA5 reanalysis archive (free, no API key): hourly temperature, shortwave radiation, and wind at the house coordinate, requested in `Europe/Stockholm` so the timestamps join the other panels directly. Tool: `backfill_weather_hourly.py`.
3. **Price** — energy-charts.info SE4 day-ahead spot (CC BY 4.0, Fraunhofer ISE), with the 2025-10-01 15-minute market change resampled to the hourly training standard, backfilled to 2022-01 so the price panel now spans the same window as consumption and weather. Tool: `backfill_se4_prices.py`.

This is **reproducible by anyone** with a Swedish electricity contract and a free weather API — no proprietary feed, no paid service, no dedicated instrumentation — which lowers the barrier to N = 1 occupied-home evidence substantially, where the living-lab literature relies on institutional facilities. It is **firewall-clean**: the panels live entirely on the research plane and never touch the frozen IPMVP baseline. Its provenance and licensing are fully traceable (Table 2).

Table 2: Evidence-panel provenance and licensing.

Source	Licence / access	Tool (committed)	Panel	Coverage
Grid "Mina sidor" hourly export	Customer's own metered data (free)	`merge_consumption_hourly.py`	Consumption (whole-house, hourly)	39,403 h, 2022-01 → 2026-06 (0.9999)
Open-Meteo ERA5 archive	Free, no key (attribution)	`backfill_weather_hourly.py`	Weather (temp/SW/wind)	39,408 h, 2022-01 → 2026-06 (1.0)
energy-charts.info SE4 day-ahead	CC BY 4.0 (Fraunhofer ISE)	`backfill_se4_prices.py`	Price (SE4 spot, hourly)	39,407 h, 2022-01 → 2026-06 (1.0)
Three-panel join	—	join on `(date_local, hour_local)`	Cons. × weather × price	**39,403 h** (3-way overlap)

All panels keyed on `(date_local, hour_local)`, localised to Europe/Stockholm before merge; raw panels are de-identified whole-house energy only (no per-appliance data, no PHI). Coverage counts are hours after localisation and de-duplication (the 0.9999 consumption fraction is ∼5 DST hours).

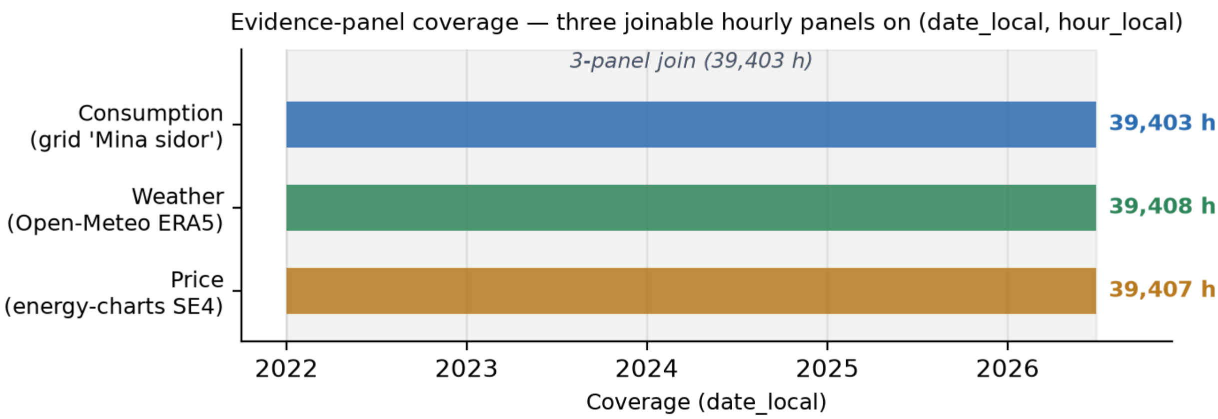

Figure 1: Coverage of the three hourly panels. All three panels span 2022-01 → 2026-06. The shaded region is the three-way joinable window (39,403 h). *F1–F3 are schematics (no underlying data); F4 and F5 are the only data-bearing figures.*

Figures

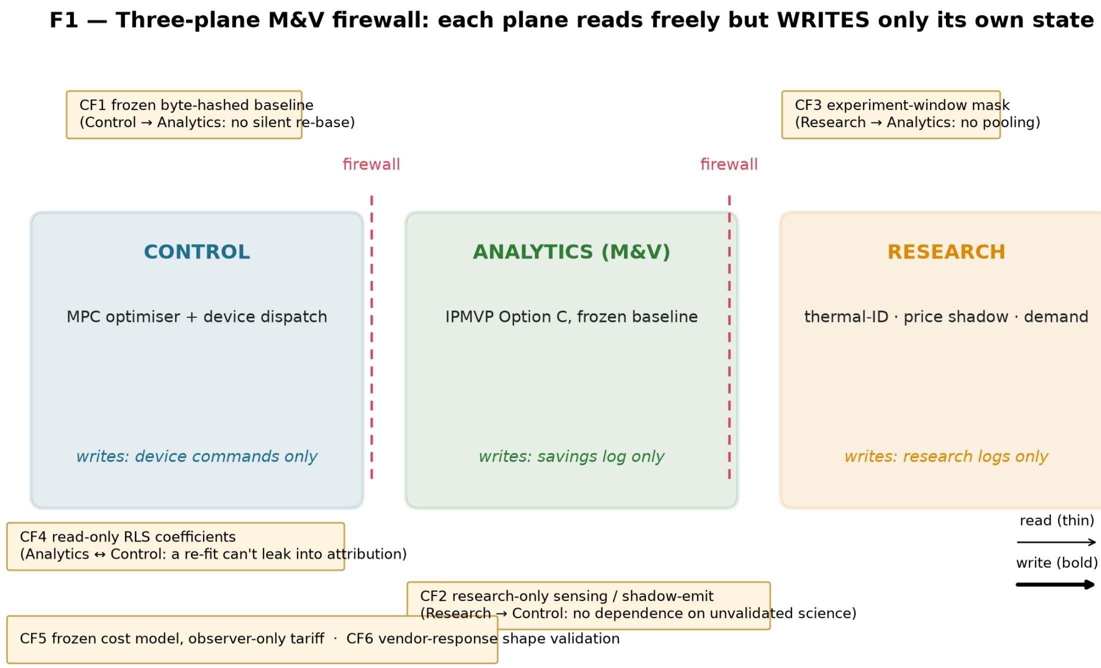

Figure 2: F1 — the three-plane M&V firewall: each plane reads freely but writes only its own state; the six FMEA guards (CF1–CF6) annotated on the cross-plane paths they close.

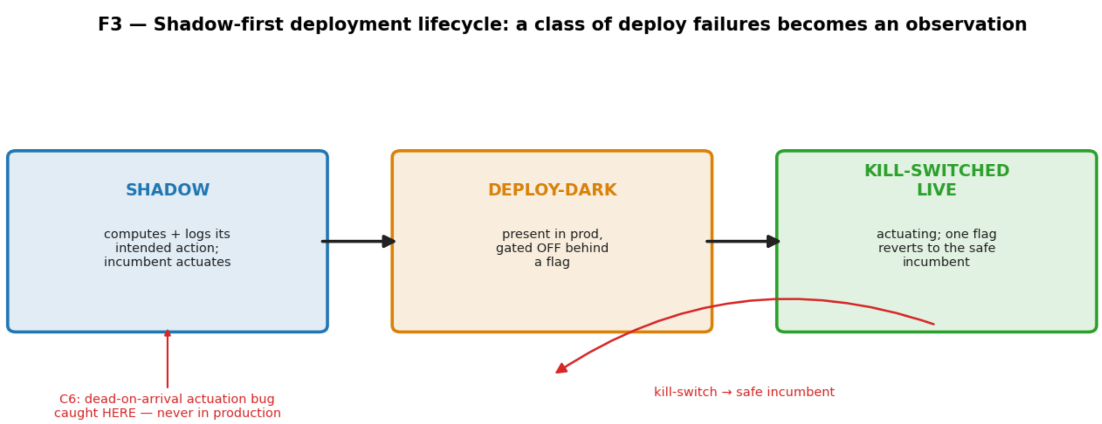

Figure 3: F3 — the shadow-first deployment lifecycle (shadow → deploy-dark → kill-switched live). The C6 dead-on-arrival actuation bug was caught in the shadow stage, never in production.

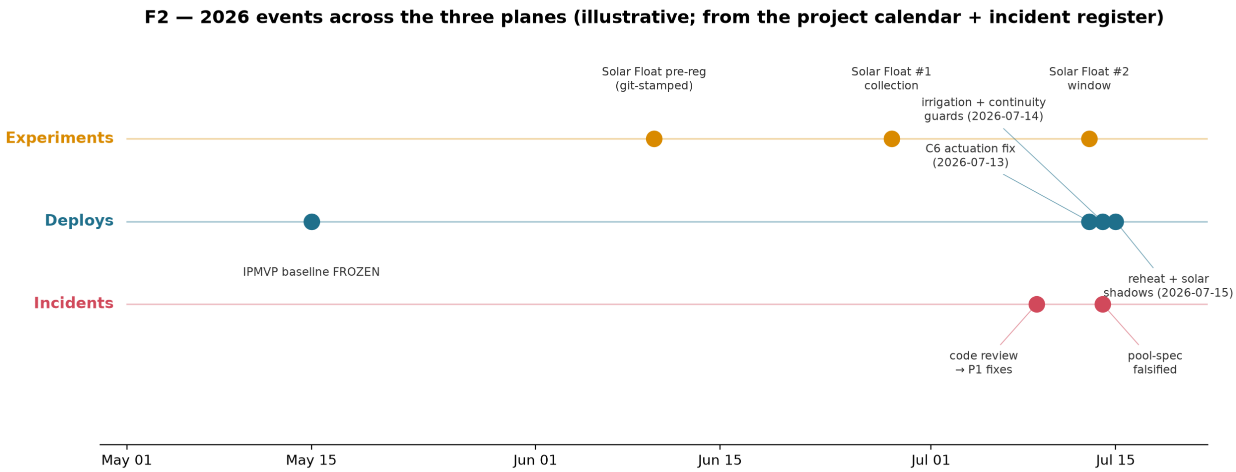

Figure 4: F2 — 2026 events across the three planes (illustrative; from the project calendar and the incident register).

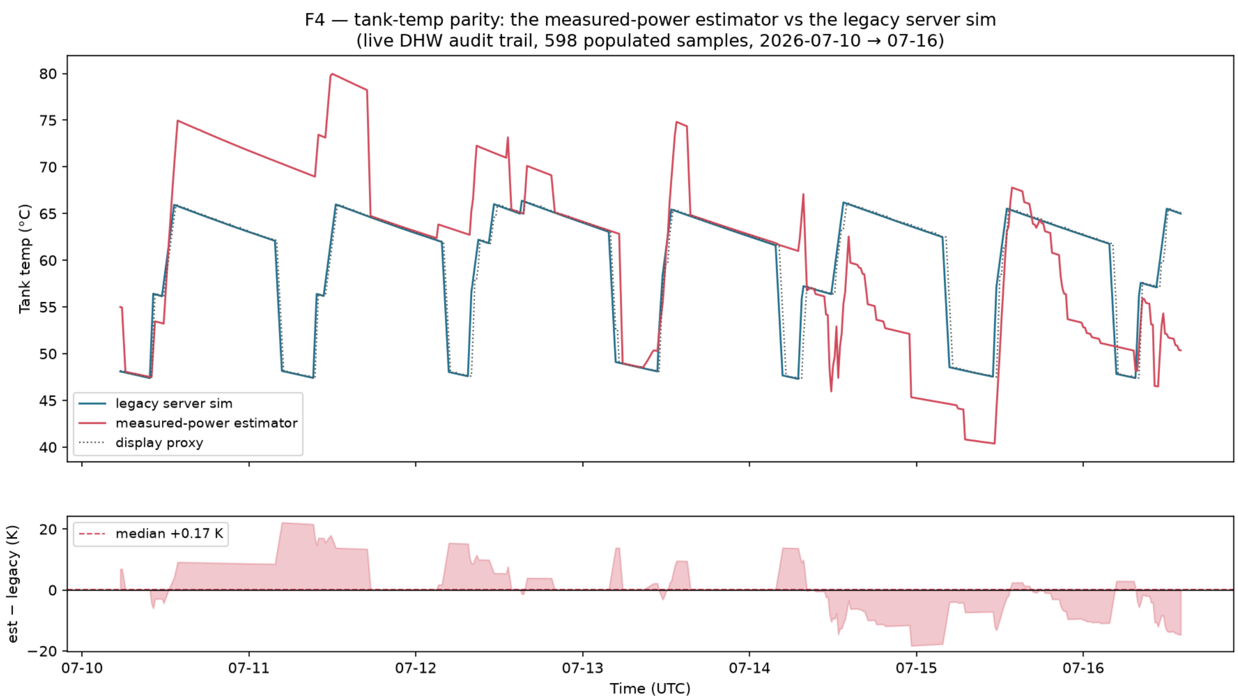

Figure 5: F4 — "drift is the measurement": the measured-power tank-temperature estimator vs the legacy server simulation, from the *live* DHW audit trail (598 estimator-populated samples, 2026-07-10→07-16). The estimator−legacy divergence band (median +0.17 K, range −18.3/ + 22.2 K) sign-flips across the window and excurses wider during charge sawtooths — a live calibration signal for the next controller revision's cutover.

Declaration of generative AI and AI-assisted technologies

During the preparation of this work the author used Claude (Anthropic) — in agentic configurations for data analysis support, figure-generation code, and manuscript drafting. All analyses were executed under a version-controlled verification discipline defined by the author: every numeric claim is traced to a source record in an auditable numbers ledger, all figures are generated from committed code paths, and drafts passed an independent traceability review before acceptance. After using these tools, the author reviewed, verified and edited all content, and takes full responsibility for the content of the publication. AI tools are not credited with authorship. (The agent-discipline methodology itself is the subject of a companion paper.)

Data & Materials Availability

Two of the internal artefacts cited as evidence are released as a **de-identified supplementary deposit** (Zenodo, CC-BY-4.0, doi:10.5281/zenodo.21420257): the three-plane M&V-firewall FMEA [3] and the free-float pre-registration record [4] — de-identified derivatives (house/zone/device/vendor names and internal cross-references generalised) that preserve the architecture and guard logic, the freeze statement, and the falsifiable hypotheses with their frozen thresholds; no personal or household data appears. The evidence-panel data (§6) are built from public or customer-owned sources with committed tooling (Table 2). The remaining internal records cited — the code review [5], the incident-register note [6], and the deploy-verification record [7] — are **available from the author as de-identified excerpts on request**. The controller source code is on a staged public-release path (2028) and is **not** required to reproduce the skills-and-gate pattern or the firewall architecture; no personal, household, or occupant-environment data is included in any released or on-request artefact.

Competing interests

The author serves on the board of Kraftringen, a Swedish municipal energy company, on a political mandate; the household studied is a retail electricity customer of Kraftringen. Kraftringen provided no funding, data, or input to this work and had no role in its design, analysis, or conclusions. The price data used are public SE4 day-ahead market prices; savings are computed against the household's own metered consumption.

References

[1] A. J. Khabbazi, E. N. Pergantis, L. D. Reyes Premer, P. Papageorgiou, A. H. Lee, J. E. Braun, G. P. Henze, K. J. Kircher. "Lessons learned from field demonstrations of model predictive control and reinforcement learning for residential and commercial HVAC: A review." *Applied Energy* 399 (2025) 126459. doi:10.1016/j.apenergy.2025.126459. arXiv:2503.05022.

[2] E. N. Pergantis, Priyadarshan, N. Al Theeb, P. Dhillon, J. P. Ore, D. Ziviani, E. A. Groll, K. J. Kircher. "Field demonstration of predictive heating control for an all-electric house in a cold climate." *Applied Energy* 360 (2024) 122820. doi:10.1016/j.apenergy.2024.122820. arXiv:2402.07032.

[3] R. Elmér. "M&V firewall — the three-plane (Control / Analytics / Research) FMEA." De-identified supplementary deposit, 2026. Zenodo, doi:10.5281/zenodo.21420257 (CC-BY-4.0), `deposit/three_plane_mv_firewall.md`.

[4] R. Elmér. "Free-float 168-hour analysis plan (pre-registration record)." De-identified supplementary deposit, 2026 (plan git-timestamped 2026-06-10, frozen before data collection). Zenodo, doi:10.5281/zenodo.21420257 (CC-BY-4.0), `deposit/preregistration_free_float.md`.

[5] R. Elmér. "Gold-standard code review of the controller (two-part, server revision r83.30/31)." Internal project review record, 2026-07-09. Available from the author as a de-identified excerpt on request (see Data & Materials Availability).

[6] R. Elmér. "A self-confirming echo loop in the controller's solar input: detection against an independent satellite series, root cause, and the loop-breaking fix." Internal incident-register note, 2026-06-25. Available from the author as a de-identified excerpt on request.

[7] R. Elmér. "Post-float deploy verification (byte-identical dev↔prod)." Internal deploy record, 2026. Available from the author as a de-identified excerpt on request.

[8] E. N. Pergantis, L. D. Reyes Premer, A. H. Lee, Priyadarshan, H. Liu, E. A. Groll, D. Ziviani, K. J. Kircher. "Protecting residential electrical panels and service through model predictive control: A field study." arXiv:2409.04884, 2024.

[9] T. Brudermueller, E. Fleisch, M. González Vayá, T. Staake. "HEAPO — an open dataset for heat pump optimization with smart electricity meter data and on-site inspection protocols." arXiv:2503.16993, 2025. (Proc. 16th ACM Int. Conf. on Future and Sustainable Energy Systems, doi:10.1145/3679240.3734637.)

[10] J. Drgoňa, J. Arroyo, I. Cupeiro Figueroa, D. Blum, K. Arendt, D. Kim, E. Perarnau Ollé, J. Oravec, M. Wetter, D. L. Vrabie, L. Helsen. "All you need to know about model predictive control for buildings." *Annual Reviews in Control* 50 (2020) 190–232. doi:10.1016/j.arcontrol.2020.09.001.

[11] D. Sturzenegger, D. Gyalistras, M. Morari, R. S. Smith. "Model predictive climate control of a Swiss office building: Implementation, results, and cost–benefit analysis." *IEEE Transactions on Control Systems Technology* 24(1) (2016) 1–12. doi:10.1109/TCST.2015.2415411.

[12] J. R. Vázquez-Canteli, Z. Nagy. "Reinforcement learning for demand response: A review of algorithms and modeling techniques." *Applied Energy* 235 (2019) 1072–1089. doi:10.1016/j.apenergy.2018.11.002.

[13] Efficiency Valuation Organization. *International Performance Measurement and Verification Protocol (IPMVP), Core Concepts.* EVO 10000-1. Efficiency Valuation Organization, 2022. https://evo-world.org.

[14] ASHRAE. *ASHRAE Guideline 14-2002: Measurement of Energy and Demand Savings.* American Society of Heating, Refrigerating and Air-Conditioning Engineers, Atlanta, 2002.

[15] J. Granderson, P. N. Price. "Development and application of a statistical methodology to evaluate the predictive accuracy of building energy baseline models." *Energy* 66 (2014) 981–990. doi:10.1016/j.energy.2014.01.074.

[16] J. Granderson, S. Touzani, C. Custodio, M. D. Sohn, D. Jump, S. Fernandes. "Accuracy of automated measurement and verification (M&V) techniques for energy savings in commercial buildings." *Applied Energy* 173 (2016) 296–308. doi:10.1016/j.apenergy.2016.04.049.

[17] S. Kapoor, A. Narayanan. "Leakage and the reproducibility crisis in machine-learning-based science." *Patterns* 4(9) (2023) 100804. doi:10.1016/j.patter.2023.100804. arXiv:2207.07048.

[18] G. M. Huebner, M. J. Fell, N. E. Watson. "Improving energy research practices: guidance for transparency, reproducibility and quality." *Buildings and Cities* 2(1) (2021) 1–20. doi:10.5334/bc.67.

[19] Lawrence Berkeley National Laboratory. *FLEXLAB: building efficiency testbed.* https://flexlab.lbl.gov (accessed 2026).

[20] A. L. Pisello, T. M. Amer, B. Pioppi. "Ten questions concerning living laboratories (LL) for multi-domain comfort investigations." *Building and Environment* 276 (2025) 112750. doi:10.1016/j.buildenv.2025.112750.

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

[22] A. Arteconi, D. Patteeuw, K. Bruninx, E. Delarue, W. D'haeseleer, L. Helsen. "Active demand response with electric heating systems: Impact of market penetration." *Applied Energy* 177 (2016) 636–648. doi:10.1016/j.apenergy.2016.05.146.

[23] U. Forssell, L. Ljung. "Closed-loop identification revisited." *Automatica* 35(7) (1999) 1215–1241. doi:10.1016/S0005-1098(99)00022-9.

[24] D. Kim, J. Cai, K. B. Ariyur, J. E. Braun. "System identification for building thermal systems under the presence of unmeasured disturbances in closed loop operation: Lumped disturbance modeling approach." *Building and Environment* 107 (2016) 169–180. doi:10.1016/j.buildenv.2016.07.007.

[25] D. Kim, J. Cai, J. E. Braun, K. B. Ariyur. "System identification for building thermal systems under the presence of unmeasured disturbances in closed loop operation: Theoretical analysis and application." *Energy and Buildings* 167 (2018) 359–369. doi:10.1016/j.enbuild.2017.12.007.

[26] X. Wang, W. Tang. "Analysis and evaluation of baseline manipulation in demand response programs." arXiv:2011.10681, 2020.