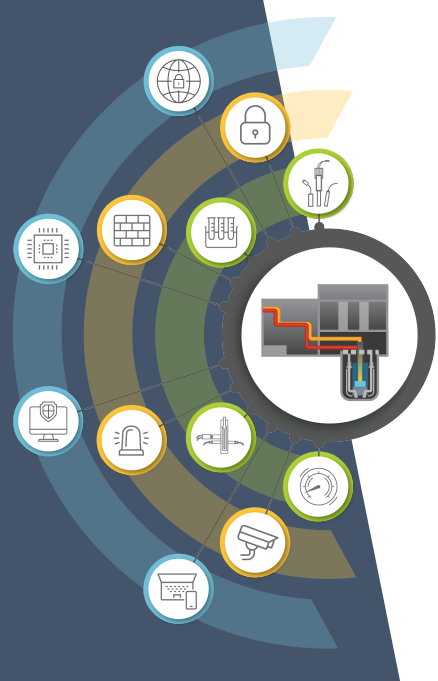

PHYSICS-BOUNDED ROBUSTNESS EVALUATION OF ANOMALY DETECTORS FOR INDUSTRIAL CONTROL SYSTEMS

PRESENTED BY

Benjamin Blakely

April 2026

APT-203540

Argonne NATIONAL LABORATORY

U.S. DEPARTMENT OF ENERGY

Primary Effort: Research Question

PI: Benjamin Blakely

Standard robustness evaluation perturbs inputs arbitrarily.
This work perturbs within *the physical system's validated constraints*.

The robustness question

Process physics bound what perturbations are physically realizable. Within those bounds, perturbations can be individually admissible yet cumulatively significant — the anomaly detector may never see a single reading that looks wrong.

Prior work evaluates detectors on obviously anomalous data. **This work** evaluates on *admissible* perturbations that respect the operating envelope.

The central hypothesis

All five primary detector classes evaluate at *one temporal scale* and *one derivative order* — structurally insensitive to locally-admissible but cumulatively-significant perturbations, regardless of tuning.

The experiment: stress each detector's scoring assumption with physics-bounded perturbations and quantify the residual detection gap.

Data Characterization

Telemetry

- 134-day window, Feb–Jun 2025
- Process historian: RO + RW sensors
- 30-second sampling interval
- 2.5B readings

Steady-state filter

- p90 stability index (sensor σ, range) per row group
- 96% of 2,472 admissible; 4% excluded (startup/shutdown transitions)

Preprocessing

- Long format: one row per sensor per timestamp
- Categorical fields label-encoded
- Values scaled with RobustScaler
- Boolean signals mapped 1.0/0.0

Train/test split: Models train on full 134-day history; attack episodes are generated separately — no overlap.

The Eight Detectors

	Detector	What it learns	Anomaly score
Primary paradigms — five distinct scoring assumptions			
A	LSTM	Next sensor values at 3 horizons (30s, 2.5min, 15min)	Prediction error
B	Autoencoder	Compressed normal-data manifold	Reconstruction error
C	Change-point	Local sensor distribution	Discontinuity magnitude
D	Dependency violation	Pairwise cross-sensor correlations	Weighted correlation breakdown
Causal ablation group — each relaxes one DYNOTEARS assumption			
E	DYNOTEARS	Linear causal DAG over sensors	Structural Hamming Distance (SHD)
F	NTS-NOTEARS	Nonlinear causal structure (1D-CNN)	SHD — tests linearity assumption
G	PCMCI+	Causal structure with hidden confounders	SHD — tests confounder assumption
H	Bayesian network	Probabilistic BN structure	Log-likelihood — tests SHD as score

Each detector embeds a different scoring assumption; the evaluation suite stresses those assumptions independently.

E–H causal ablation: decomposes whether DYNOTEARS is limited by linearity, confounders, or the choice of SHD as the anomaly score.

Physics Validator

ColdTrapValidator — 10 constraints

- Absolute temperature bounds (100–300°C)
- Energy balance: ΔT preservation
- Flow-temperature polynomial (empirical)
- Spatial uniformity ($<$5°C across sensors)
- Per-sensor rate-of-change limits
- Regime-aware: IDLE / STARTUP / ACTIVE / SHUTDOWN

These constraints define the boundary of *physically realizable* perturbations.

Two evaluation modes

Bounded	Physics on Perturbations clipped
Unbounded	Physics off Unconstrained

Bounded vs. unbounded isolates whether a detector relies on constraint violations or detects genuine behavioral anomalies. A detector that works only in unbounded mode fails against careful perturbations.

Admissibility Ladder

Severity graded by **beta distributions** fitted from empirical sensor statistics and physics constraints.

Tier	Class	Beta shape (α, β)	Support	Definition
1	Empirically admissible	$\gg 1$ (concentrated bell near mean)	$[p_1, p_{99}]$	Within observed variation
2	Physics-consistent	$\approx$ 3–5 (broad bell)	Physics bounds	Exceeds history; physics satisfied
3	Stress-case	< 1 (U-shaped, mass at bounds)	Physics bounds	Approaches constraint limits
4	Physics-violating	< 1, $\pm 10\sigma$ extension	Beyond physics	Positive control — must be detected

Key question: At which tier does each detector fail? A detector catching only Tier 4 provides no operational value.

Detector Evaluation Methodology

For each detector × anomaly signature × tier × mode:

1. **Train once** on the full 134-day clean dataset. Cached and reused across all 63 anomaly signature evaluations.
2. **Generate 50 perturbation episodes:** random 1-hour windows with the perturbation signature applied. *These are the independent experimental units.*
3. **Score perturbed data:** Slide a 25-min window across each episode. Each episode yields ∼70 overlapping windows; 3500 total scored sequences. Score = prediction error (LSTM), reconstruction error (AE), change magnitude (CPD), correlation breakdown (DV), or SHD (causal).
4. **Score baseline:** Same model on ∼2000 normal sequences from 8 evenly-spaced segments of the 134-day history (restricted to steady-state row groups — see Data Characterization).
5. **Compare distributions** via ROC analysis (next slide).

ROC Analysis Methodology

ROC construction

- Normal sequences $y = 0$, perturbed sequences $y = 1$
- Sweep threshold: compute TPR and FPR
- **AUC** = P(perturbed score $>$ normal score)
  Threshold-independent measure of score separability

Reported KPIs per detector $\times$ anomaly signature

AUC	Score separability
p (n=50 episodes)	Statistical significance
TPR @ 1% FPR	Tight operational point
TPR @ 5% FPR	Relaxed operational point

Why episode-level?

3500 sequences are *not* independent: consecutive 25-min windows share $\sim$96% of data.

Episode-level (n=50, paper):
Count pairs where perturbed score $>$ normal score; test whether win rate $\gg$ chance.

Sequence-level ($n\approx$3500, reference):
Overstates 7$\times$; both reported.

AUC$\approx$0.98 $\Rightarrow$ $p \ll$0.001 either way.

Current Work in Progress

Running now (April 2026)

- Attack generation: 16 anomaly signatures × 4 tiers × 2 modes, admissibility-filtered
- 4 compute hosts parallelizing across ~400 episodes per signature
- Preprocessing and detector training queued

Experiment matrix (May 2026)

- 120 primary experiments (5 detectors × 3 regimes × 4 tiers × 2 modes)
- 36 causal ablation runs (gated on DYNOTEARS)

Planned analyses

- AUC heatmap: anomaly signature × detector
- Bounded vs. unbounded: isolates physics-exploiting detectors
- Tier monotonicity: detection rate vs. perturbation severity
- Causal ablation: linearity vs. confounders vs. likelihood scoring

Parallel Effort: Multi-Modal Collection

PI: Benjamin Blakely
Purpose: Real-time, streaming capture of ICS-adjacent network and operational data from the source facility to Argonne datacenter storage.

Architecture

- Source-facility head-end collection node → Apache Arrow Flight → Argonne datacenter storage node
- Three parallel capture streams with *cross-stream activation*: any event raises collection frequency on all streams simultaneously
- Partitioned Parquet storage (facility / stream / date)

Data modalities

- **Telemetry:** process historian (ISA-95 schema, engineering units, quality flags)
- **NetFlow v5:** network traffic flows, bytes/packets, protocol, routing
- **IDS alerts:** Suricata EVE-JSON (signatures, severity, DNS/HTTP/TLS metadata)

Delivered components

- `arss_collector` — containerized multi-stream collection agent (Docker Compose), deployed at source facility
- `arss_collector_server` — central Flight server at Argonne; token-authenticated, streaming ingestion
- `arss_collector_library` — Python query library (Polars-based, ~160 tests passing); pip-installable

Multi-Modal Collection: Provided to SRNL

Delivery timeline

- **Dec 2025** — Director's discretionary allocation on Argonne HPC (Polaris) obtained; access credentials provided to SRNL
- **Jan 2026** — Server and network infrastructure installed, configured, and operational; data collection active
- **Feb 4, 2026** — On-site meeting with SRNL (Glenn Fink, Argonne) to demonstrate system and discuss next steps
- **Feb 2026** — Topology documentation, `arss_collector_library`, and initial network data tranche delivered to SRNL

Working with facility team to obtain more comprehensive data coverage; multiple complete capture windows already provided to SRNL for initial testing.

Current status

- Infrastructure remains installed and operational at the source facility
- All code, data, and HPC access delivered to SRNL as of Feb 2026
- Pending confirmation of SRNL utilization of code, data, and Polaris allocation

Argonne deliverables summary

- ✓ Capture infrastructure (operational)
- ✓ Data access library (tested, packaged)
- ✓ HPC allocation + credentials
- ✓ Topology documentation
- ✓ Initial data tranche
- ✓ On-site technical walkthrough

Parallel Effort: Physics-Informed Anomaly Detection

PI: Rick Vilim

Completed (Jan–Feb 2026)

- **Hyperplane detector:** normal-operation hyperplane constructed for steady-state; deviations flagged as candidate attacks
- **Neural ODE simulator:** models cold-trap transient behavior; packaged as FMU (denoiser → normalizer → Neural ODE → denormalizer) for online forecasting
- Simulator evaluated against multiple attack scenarios including stealthy attacks; simulator uncertainty under adversarial conditions quantified

Planned methodology

- **Physics-embedded Neural ODE:** embed physical constraints directly into model architecture; compare against other physics-informed techniques
- **Residual analysis:** monitor deviations between Neural ODE predictions and observed system behavior as attack signatures
- New metrics to quantify detection capability and characterize residual signals under normal vs. abnormal conditions
- Integration into online monitoring pipeline for real-time applicability

Physics-Informed Anomaly Detection: Outlook

Deliverables to date

- ✓ Steady-state hyperplane detector
- ✓ Neural ODE simulator (FMU-packaged)
- ✓ Adversarial uncertainty characterization

Planned upon resumption

- Physics-embedded Neural ODE development and testing
- Comparison with existing physics-informed techniques; performance evaluation across operating conditions
- Residual signal characterization under normal and abnormal conditions
- Online pipeline integration for real-time cyber-attack detection

Future Work

Phase 3 — Statistical rigor (June 2026)

- 5-seed repeats for confidence intervals and significance testing
- Continuous robustness-boundary sweeps: *P*(detection) as a function of perturbation magnitude
- Ablation studies isolating each detector modification's contribution

Phase 4 — Dissemination (Jul–Aug 2026)

- ANS Winter 2026: admissibility envelope methodology (Jun 12 deadline)
- Technical report

FY27 — Residual indicators of compromise

- Characterize sub-threshold behavioral artifacts that persist when detectors fail to alert
- Physics-grounded ICS IOC taxonomy derived from the admissibility ladder