DIGITAL TWIN WHITE PAPER SERIES

ENERGY & UTILITIES · EDITION 02 · AUGUST 2026

The Number You Cannot Audit



Predictive maintenance and digital twins in power generation — what the published results actually show, and how to tell a defensible KPI from a marketing one.

Lise DELOBEL

GenAI project manager & consultant — digital twins, Physical AI and applied ML for industrial operations. MIT IDSS Data Science & Machine Learning certificate.

LISEDELOBEL.FR · LISE.DELOBEL@ME.COM

Strong claims, thin evidence, real value

The argument in one line: digital twins can create measurable maintenance value in power generation, but the strongest business case rests on observable performance, fault-detection evidence and explicit counterfactual assumptions — not on unattributed savings percentages.

Commercial literature repeatedly cites figures that are striking and consistent in shape: unplanned downtime down 20 to 50 percent, maintenance costs down 15 to 25 percent against time-based strategies, thermal efficiency up 0.4 to 2.5 points. For a 500 MW combined-cycle plant those ranges are quoted at 1.2 to 3.8 million dollars a year. Every one of them is reported rather than independently validated, and none of the sources publish the arithmetic behind them — which, by the standard argued for here, already disqualifies the number from a board pack.

Meanwhile a 2025 academic survey of digital twins in large-scale power plants concludes there is a clear research gap around the practical constraints of implementing them. Both things hold at once: the technology does work, and the evidence base for how well is weaker than the marketing suggests. The executive skill that matters is therefore not platform selection. It is knowing which KPI will bear weight.

FINDING 01

The value is real where the physics is legible

The strongest evidence is a mechanism rather than a percentage. Industry technical literature attributes a large share of recoverable compressor-related deterioration — commonly cited around 70 percent — to fouling, and one heavy-duty reference reports a 1 point loss of compressor efficiency corresponding to roughly 1.1 percent of output. Where degradation has a known physical signature, a twin has something to find.

FINDING 02

Most circulating KPIs cannot be reproduced

Downtime and cost ranges are typically self-reported by platform vendors, span a factor of two or more, and rest on a counterfactual — the failure that did not happen — which was never observed. That does not make them false. It makes them irreproducible by a third party, which is a different and more practical problem.

FINDING 03

Avoided cost is model-dependent, not unknowable

A prevented trip valued at several hundred thousand dollars assumes the trip would have occurred and that the intervention prevented it. Such benefits can be estimated rigorously — with base rates, matched assets or survival models — but the result is auditable only to the extent that those assumptions, probabilities and attribution rules are themselves auditable.

FINDING 04

Three outcome KPIs survive scrutiny

Detection lead time on confirmed faults; alert precision paired with confirmed-fault recall; and corrected thermodynamic performance against a clean-condition baseline. A fourth of a different kind — validated asset coverage — measures the programme rather than the outcome. Few appear in business cases.

Two definitions used throughout

In this paper, digital twin means an operational, physics-informed or hybrid asset model synchronised with operational data and used to compare expected against observed behaviour — not a visualisation and not a dashboard. Industry usage is broader. Detection lead time is the interval between a qualifying alert and independent confirmation of the same failure mode — observable without estimating an avoided failure.

Four pressures converging on the plant

Digital twins have been discussed in power generation for a decade. Four things have changed what they are for.

~70%
OF RECOVERABLE LOSS
Share of recoverable compressor-related deterioration commonly attributed to fouling

1.0 → 1.1
EFFICIENCY TO OUTPUT
Points of output per point of compressor efficiency, one heavy-duty reference

$2.1bn
MARKET, 2025
Reported size of the digital twin power plant market — an analyst estimate, not a result

01 The fleet is old and the duty cycle changed

Thermal plant built for baseload now cycles to follow renewables. Start-stop duty accelerates thermal fatigue in exactly the components whose inspection intervals were set for steady running, which invalidates calendar-based maintenance assumptions without invalidating the equipment.

02 Instrumentation finally reaches the model

Historians, plus retrofitted vibration and thermal sensing, now supply the continuous state a twin needs. The constraint has moved from data availability to data quality — tag mapping, sensor drift, calibration and unlabelled outages are the practical blockers.

03 The economics are asymmetric

A forced outage on a large unit is quoted in the industry literature at hundreds of thousands to millions in lost revenue and emergency repair. Against that, the twin does not need to be right often — it needs to be right about the expensive failures, which is a different engineering target from overall accuracy.

04 Procurement has outrun evidence

Platform marketing has moved faster than the peer-reviewed literature. A 2025 survey of large-scale power plant digital twins explicitly identifies a research gap in practical implementation constraints. That gap is where buyers currently make decisions, which is precisely why measurement discipline matters now.

Sources: Combined Cycle Journal outage handbook · Power Services Group technical notes · survey of digital twin implementation in large-scale power plants, 2025 · commercial analyst market compilations. Market and savings figures are analyst or vendor estimates, reported and not independently validated.

Measured, observed, estimated, counterfactual

Every predictive maintenance business case rests on one of these four evidence classes. They are not equally defensible, and the most quoted is the weakest.

CLASS 01

MEASURED — directly instrumented

Corrected thermodynamic performance against a clean-condition baseline: heat rate, compressor efficiency, output at reference conditions. Computed from instrumentation, comparable across vendors, and independent of anyone's model being right — provided sensor accuracy and drift are verified.

CLASS 02

OBSERVED — recorded events

Detection lead time on confirmed faults, alert precision, and confirmed-fault recall. All require only that you log what was flagged and what was later found. Recall matters as much as precision: a system that alerts rarely posts an excellent false-positive rate while missing most of the faults.

CLASS 03

ESTIMATED — before and after

Downtime hours or maintenance spend against a prior period. Legitimate, but confounded by everything else that changed — duty cycle, fuel, staffing, weather, the overhaul that happened anyway. Usable when the confounders are stated; misleading when they are not.

CLASS 04

COUNTERFACTUAL — model-dependent

The failure that did not happen, valued at what it would have cost. The avoided event cannot be directly observed, so the figure is auditable only to the extent that its base rates, probabilities and attribution rules are auditable. Report it with those exposed, never as primary justification.

THE MATURITY LADDER

01 Digital model	02 Digital shadow	03 Digital twin	04 Predictive twin	05 Prescriptive twin
A geometric or process model. No live data. Useful for design, not for maintenance.	Live data flows in, one way. You can see the asset's state; the model does not act.	Operationally integrated: data update the model, and model outputs inform decisions.	Residuals are projected forward to estimate remaining life and intervention timing.	The twin schedules the intervention against cost, outage windows and market price.

The strategic read

Many programmes marketed as digital twins are closer to digital shadows, while most of the value claimed for them belongs to rungs four and five. The distinction is not pedantic: a shadow tells you what is happening, which you largely knew; a twin tells you what should be happening and prices the difference. When evaluating a proposal, the question that separates them is simple — what does the model predict, and what happens when the plant disagrees with it?

The degradation you can price

Gas turbine compressor fouling — the best-evidenced predictive maintenance case in the plant

If you want one application where the physics, the economics and the measurement all line up, this is it. Compressor fouling is progressive, thermodynamically visible, financially significant, and one of the principal recoverable degradation mechanisms — a rare combination, and the reason credible programmes start here.

THE PHYSICS

The mechanism

Particles small enough to pass the inlet filtration deposit on compressor blade surfaces. Surface roughness rises, the boundary layer thickens, and the designed aerofoil profile is compromised. Mass flow and pressure ratio both fall; the control system compensates with fuel to hold exhaust temperature, so heat rate rises while output falls at the same time.

THE NUMBERS

The magnitude

Industry technical sources commonly attribute around 70 percent of recoverable gas turbine performance deterioration to compressor fouling, with wider estimates to 85 percent. One heavy-duty axial reference reports a 1.0 point loss of compressor efficiency corresponding to roughly 1.1 percent of output under its stated assumptions. A modelling study of a 226 MW engine used 102 micrometres of blade roughness as an approximation of about twelve months of fouling exposure, and found design-point power reduced by 8.7 percent.

THE TRADE

The recovery, and its cost

Washing restores much of it. A published evaluation on one engine reports about 2.28 percent of output regained by online washing, 5.85 percent by offline, and 6.50 percent by both combined, against roughly 7.20 percent lost to fouling. But offline washing requires shutting the machine down, so recovery is bought with lost generation — turning maintenance into a scheduling problem rather than a technical one.

So what

This is the case to build first, because it can be evaluated without believing anyone's counterfactual: compute corrected output and heat rate against a post-wash baseline and watch the curve. Many of the required variables already exist in modern turbine-control and historian systems, though sensor quality and derived-variable accuracy must be verified first. If a proposed programme cannot demonstrate value on the best-instrumented degradation mode in the plant, it is unlikely to demonstrate it on anything harder.

Sources: Combined Cycle Journal outage handbook · Power Services Group · published compressor washing evaluation studies · gas turbine degradation modelling literature. Percentages are engine-, site- and filtration-specific; treat as indicative rather than as operating rules.

What the reported results actually say

The circulating KPIs, their provenance, and how much weight each will bear

The figures below are the ones you will be shown. They are reproduced as reported, with provenance attached, because provenance is the part that usually goes missing between a vendor deck and a board pack.

REPORTED, NOT VALIDATED

The headline ranges, as reported

Commercial and analyst sources report unplanned downtime falling 20 to 50 percent, maintenance costs falling 15 to 25 percent against time-based strategies and 40 to 60 percent against reactive ones, and thermal efficiency improving 0.4 to 2.5 percentage points. For a 500 MW combined-cycle unit these are quoted at 1.2 to 3.8 million dollars annually. None is independently validated.

THE TEST APPLIED

Why you cannot reproduce the dollar figure

To rebuild that 1.2 to 3.8 million from first principles you would need capacity factor, heat rate, fuel price, forced-outage hours, the maintenance budget baseline and the assumed improvement. None of the sources publish them. A number whose arithmetic cannot be reconstructed from stated assumptions should not carry a decision — which is the whole argument of this paper applied to its own evidence.

THE SPECIFICS

The named deployments that do exist

Vattenfall reports using digital twins across its wind fleet to reduce physical inspections and identify design changes reducing steel use in turbine construction. Siemens and NVIDIA have reported physics-informed twins reducing corrosion-related downtime in combined-cycle plants by up to 10 percent. Narrower, more specific, more credible — and correspondingly less quoted.

So what

Treat aggregate ranges as evidence of the magnitude vendors and operators report, not as evidence of expected performance at your plant. In a vendor conversation the useful move is to ask for one deployment on equipment resembling yours, with the baseline definition written down, and to accept a smaller honest number over a larger unattributed one. A vendor who can supply the first is telling you something about how they operate, not just about their product.

Sources: commercial analyst compilations of utility deployments, 2026 · power-technology.com case study review, 2025 · European Journal of Energy Research, 2025. All performance figures in this section are vendor-, operator- or analyst-reported.

04 / USE CASE 03

A KPI set that survives an audit

What to instrument, how to define it, and what to stop reporting

If the aggregate ranges will not bear weight, something has to. Each metric below is specified with a numerator, a denominator and a timestamp rule, because a KPI without those is an opinion.

THE CORE SET

The outcome metrics

Detection lead time: timestamp of independent confirmation minus timestamp of the first qualifying alert for the same failure mode. Alert precision: confirmed actionable alerts divided by total alerts. Confirmed-fault recall: the share of independently confirmed target faults for which a qualifying alert arrived inside the minimum useful intervention window — without which a vendor can post perfect precision by alerting almost never.

THE LEADING INDICATORS

The health metrics

False-alert burden: non-actionable alerts per asset-month, since precision alone hides an alert volume problem. Normalised residual error under healthy conditions, or the share of operating time the model stays inside its validated error envelope — this tells you whether the twin is still trustworthy before any fault occurs. Validated asset coverage: the criticality-weighted share of assets whose models have passed defined validation criteria and remain in calibration.

THE DEMOTIONS

What to stop reporting as primary

Avoided-cost totals, cumulative savings since inception, and percentage improvements without a stated baseline period. None are illegitimate; all are irreproducible as headline numbers, and leading with them makes a real programme look like a marketing exercise to exactly the technical audience whose support it needs.

So what

Specify these in the contract, not after go-live. Detection lead time, precision and recall are cheap to instrument on day one and close to impossible to reconstruct later, because nobody logs the alerts they dismissed. A programme that can report them after twelve months has something defensible; one that can only report avoided cost has a story.

Composite framework: constructed for this paper from measurement practice described in the reviewed literature and from the author's consulting experience. Presented as a proposed standard, not an established one.

The fault you cannot open the machine to see

Compressor fouling in a running gas turbine — inference in place of inspection

THE CHALLENGE

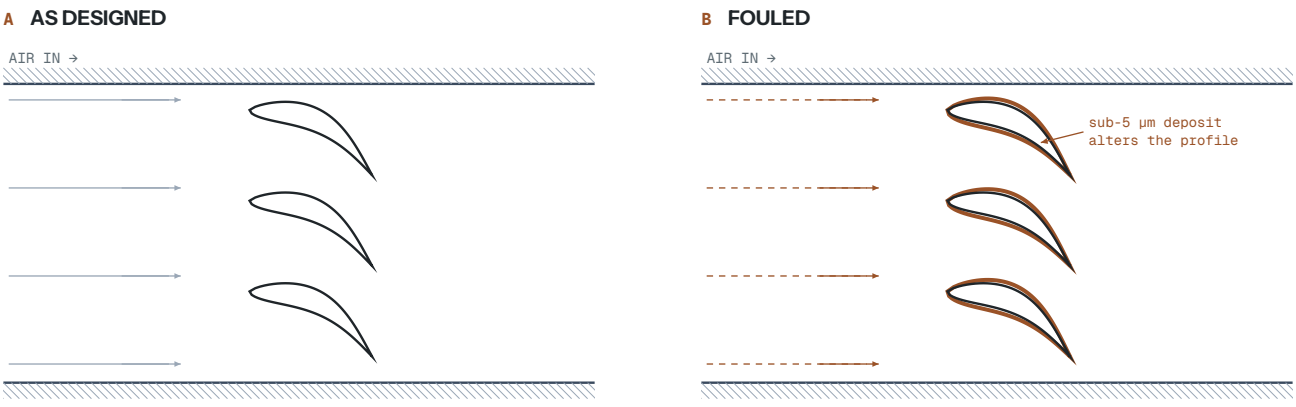
A gas turbine compressor is a stack of precisely shaped aerofoils turning at thousands of revolutions per minute inside a sealed casing. Particles small enough to pass the inlet filtration settle on those blades and change their shape. Not dramatically — a film measured in micrometres — but the aerofoil profile is the machine, and altering it alters everything downstream.

The consequences compound in a way that hides them. Surface roughness thickens the boundary layer, mass flow falls, pressure ratio falls. The control system responds by adding fuel to hold exhaust temperature at setpoint. So the unit continues to meet its dispatch instruction, the operator sees no alarm, and the loss appears only as output that quietly is not there and fuel that quietly is.

Why the obvious solutions fail

You cannot look. Borescope inspection requires the machine shut down and cooled, which costs the generation the inspection is meant to protect — and by the time you can see the deposit you have already paid for it. Calendar-based washing fails differently: fouling rate depends on ambient conditions, site dust loading and duty cycle, and is reported to vary between sites by as much as a factor of three. A fixed schedule therefore guarantees either over-washing, which wastes outage hours, or under-washing, which wastes fuel.

FIG. 01 — COMPRESSOR BLADE SECTION, CLEAN AND FOULED (SCHEMATIC, EXAGGERATED)



1 Filtered air still carries sub-5 µm particles
2 Deposits thicken the boundary layer
3 Mass flow and pressure ratio both fall
4 Control adds fuel to hold exhaust temperature
5 Output falls while heat rate rises
6 None of it is visible while running

Measure the shadow it casts

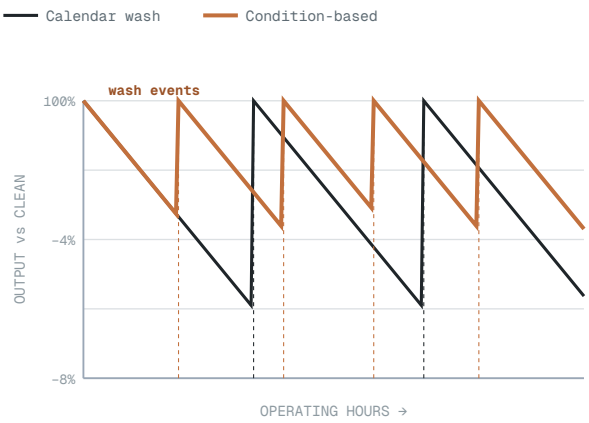
The design move, what the residual can and cannot tell you, and the trade accepted

THE RESOLUTION

The turning point is to stop trying to observe the deposit and start observing what it does. Fouling has a thermodynamic signature: for a given ambient temperature, pressure and load, a clean compressor delivers a known mass flow, pressure ratio and discharge temperature. A physics-informed model computes what those should be right now, the plant reports what they actually are, and the gap between them is measured continuously from installed instrumentation, without opening anything.

That gap is not the fouling. It is the observable signature from which fouling severity can be inferred, and only once competing causes have been accounted for — sensor bias and drift, inlet distortion, bleed configuration, variable guide vane behaviour, ambient correction error, irreversible hardware deterioration, control logic changes and plain model mismatch all produce residuals too. Attribution is the engineering work; detection is the easy part.

FIG. 02 — WASH SCHEDULING



REPORTED FIGURES — PUBLISHED STUDIES

Share of recoverable loss	~70%
Efficiency to output	1.0 → 1.1
Output lost to fouling	~7.2%
Regained, online wash	~2.3%
Regained, offline wash	~5.9%
Regained, combined	~6.5%

Recovery percentages from a published compressor-washing evaluation on one engine; share and ratio figures from industry technical sources under their stated assumptions. Values vary with engine class, filtration and site. Fig. 02 illustrates the scheduling mechanism; it is not plotted from operating data.

The trade-off accepted

The inference is only as good as its reference. A physics-informed baseline drifts as the machine genuinely ages, so fouling and irreversible wear entangle and must be separated by assumption rather than measurement — which is why the post-wash reading matters so much: it is the only moment the model gets a clean fix on the machine. Accept too that this converts a maintenance decision into a market decision, moving it across an organisational boundary most plants have not designed for.

THE OUTCOME — WHY THIS IS THE CASE TO BUILD FIRST

~70%
OF RECOVERABLE LOSS
One degradation mode carries most of the recoverable performance deterioration

8.7%
AT 102 µm ROUGHNESS
Design-point power loss in a modelling study of a 226 MW engine

0
COUNTERFACTUALS NEEDED
Every figure here is computed from installed instrumentation against a post-wash baseline

Sources: published compressor washing evaluations and gas turbine degradation modelling studies · Combined Cycle Journal · Power Services Group · US Patent 7801660 on estimating compressor fouling impact and optimising offline wash scheduling. Drawing is schematic with deposit thickness exaggerated for legibility.

Building a programme that can prove itself

Durations are indicative and depend on historian quality, asset count and outage calendar.

PHASE 00 · BEFORE ANYTHING

Write the evaluation protocol first

Define target failure modes, metrics with their numerators and denominators, attribution rules, minimum useful lead time, baseline period and counterfactual method — before the vendor sees any production results. This single step prevents the commonest evidential failure in the field, which is selecting the KPI after seeing which one looks best.

PHASE 01 · WEEKS 0-8

Fix the baseline and quantify its uncertainty

Establish corrected-performance baselines at reference conditions, and audit the historian: tag mapping, sensor accuracy, drift limits, unlabelled outages. Propagate measurement uncertainty explicitly — a claimed 0.4 percent efficiency gain means nothing if the instrumentation cannot resolve it.

PHASE 02 · MONTHS 2-6

Start where the physics is legible

Deploy on compressor fouling or an equivalent well-instrumented, well-understood degradation mode. Log every alert and its disposition from day one — including dismissed ones — because precision, recall and lead time cannot be reconstructed afterwards.

PHASE 03 · MONTHS 6-12

Extend only to modes with demonstrated precursors

Add asset classes where a physically interpretable signature exists — vibration, electrical signature, oil debris, partial discharge, thermal trend. Purely abrupt failures without measurable precursors are poor candidates, and extending to them produces disappointment that gets attributed to the technology.

PHASE 04 · 12 MONTHS +

Connect maintenance to dispatch

Move from detection to scheduling: price interventions against outage windows, forward power prices and fuel cost. The largest value sits here, and so does the largest organisational obstacle, because it requires maintenance and trading to agree on a shared objective function.

> **Organisational readiness**
>
> The characteristic failure is not technical. It is that the twin raises a recommendation nobody has authority to act on inside the window that makes it valuable. Decide before go-live who can move an outage on a model's advice, and what evidence they need — that decision right, not the model, is what converts a detection into a saving.

What goes wrong, and what it costs

Ranked by how often it is missed rather than by severity.

RISK 01

Irreproducible headline KPIs

Avoided-cost totals rest on a counterfactual never observed and assumptions rarely published. Fine as supporting colour, dangerous as primary justification before a technical audience who will spot it immediately.

RISK 02

Measurement uncertainty ignored

A twin cannot demonstrate a 0.5 percent efficiency gain when the combined uncertainty of sensors and corrections is of the same order. Uncertainty must be propagated and stated, not assumed away.

RISK 03

Baseline drift

A residual is meaningless without a valid reference. As the machine genuinely ages, fouling and irreversible wear entangle, and a model still comparing against commissioning conditions reports degradation no wash will recover.

RISK 04

Maintenance-selection bias

Teams inspect the alerts they judge credible and ignore the rest, which biases the labels and makes later performance statistics look considerably better than the system actually is.

RISK 05

Alert fatigue

False positives are the commonest cause of abandonment and the metric nobody instruments. Once operators dismiss flags by habit the true positives go with them, and the failure is invisible in the reporting.

RISK 06

Concept drift

Fuel composition, dispatch regime, upgrades, control-logic changes and component replacements all invalidate previously calibrated residual thresholds. A model validated once is not validated permanently.

RISK 07

Shadow sold as twin

Many deployments visualise live data without predicting anything. Useful, but they cannot deliver the value case that justified them. Ask what the model predicts and what happens when the plant disagrees.

RISK 08

Data quality discovered late

Tag mapping errors, drifting sensors and unlabelled outages surface months in and are usually the real cause of poor model performance. Cheap to audit up front, expensive to diagnose from inside a model.

RISK 09

Wrong failure mode

Residual-based detection needs a physically interpretable precursor. Applied to abrupt stochastic failures it finds little, and the disappointment is attributed to the technology rather than to the selection.

RISK 10

No authority to act

A correct recommendation that cannot be executed inside the valuable window produces documentation, not savings. This is an organisational design question routinely treated as an implementation detail.

Seven questions for any value claim

01 What exactly was measured?

02 What was inferred rather than measured?

03 What is the baseline, and over what period?

04 What counterfactual is assumed, and at what base rate?

05 What denominator does each rate use?

06 What uncertainty surrounds the number?

07 Could a third party reproduce it from retained plant data?

Apply it to your own reporting before applying it to a vendor's. The last question is the one that decides the others.

NEXT IN THIS SERIES

Edition 03 — Three Lots, One Plant

Inside a waste-to-energy build: the electrical scope across turbo-alternator, fire infrastructure and plant automation — and why system customisation, not installation, set the critical path.

SOURCES & METHOD

Method note: this edition separates evidence classes deliberately. Fouling mechanisms, efficiency-to-output ratios and washing recovery percentages come from industry technical publications and published engineering evaluations, under their stated engine and site assumptions. Aggregate downtime, cost and efficiency ranges are vendor-, operator- or analyst-reported, are not independently validated, and are identified as such wherever they appear. The KPI framework in section 04 and the audit test in section 08 are proposed standards constructed for this paper.

Technical: Combined Cycle Journal outage handbook · Power Services Group on compressor degradation · published compressor water-wash evaluation studies · gas turbine degradation modelling literature · US Patent 7801660.

Deployment and market: power-technology.com utility case study review, 2025 · European Journal of Energy Research, 2025 · survey of digital twin implementation in large-scale power plants, 2025 · commercial analyst market compilations, 2026.

Lise DELOBEL

GenAI project manager & consultant — digital twins, Physical AI and applied ML for industrial operations.
MIT IDSS Data Science & Machine Learning certificate.

LISEDELOBEL.FR · LISE.DELOBEL@ME.COM

DIGITAL TWIN SERIES · ENERGY & UTILITIES · EDITION 02 · AUGUST 2026