A Cross-Phase Prognostic Framework for Emergent Work Risk Assessment and Schedule Decision Support in Nuclear Refueling Outages

D. Mandelli[1] and C. Wang[1]

[1] *Idaho National Laboratory, Idaho Falls, ID 83415, USA*
diego.mandelli@inl.gov
congjian.wang@inl.gov

ABSTRACT

Refueling outages are among the most schedule-critical phases in a nuclear plant's operating life. The premise of this paper is that outages are governed by two persistent challenges across both the planning and execution phases. The first is the inherent variability in maintenance activity completion times. Activity durations depend on component health state, crew availability, and field findings that deviate from the plan, yet current practice sizes contingency buffers through engineering judgment rather than data-driven uncertainty quantification. The second is unplanned activities emerging during outage execution, where a situation found during a maintenance, testing, or surveillance activity (with potential consequences on outage execution) requires to be addressed. This paper presents a cross-phase analytical framework that addresses both of these challenges using a single architecture built on historical condition report records, work order histories, and outage schedule data as primary inputs. In the planning phase, a proactive pipeline processes the multi-cycle condition report and work order history through a sequence of analytical stages that construct a component knowledge graph, assess component health-state trajectories by analyzing degradation trends across successive outage cycles, and compute a composite risk score combining degradation frequency, prior emergent work precedent, and component criticality. Components exhibiting escalating health deterioration are flagged even when no prior unplanned maintenance record exists, enabling the detection of first-failure components that would otherwise receive a zero risk score under conventional evidence-based methods. In the execution phase, a reactive pipeline accepts a condition report, and it produces a structured recommendation (proceed, defer, escalate, or monitor) supported by a duration estimate derived from the semantic retrieval of analogous historical work orders at the median and 80th-percentile levels via relevance-weighted empirical quantiles, alongside an analog count as a calibrated confidence signal.

1. INTRODUCTION

Refueling outages of nuclear power plants are complex phases where thousands of activities are completed in just a few weeks. These phases have a major impact on plant finances since every additional hour of outage delay can cost hundreds of thousands of dollars. Sources of completion delays are mainly of two types: emergent (i.e., unplanned) activities that need to be added to the outage work scope, and delays in completing planned activities.

Emergent work (e.g., maintenance discovered mid-outage when equipment is opened for inspection) cannot be avoided: wear, packing leakage, and instrumentation drift appear only during the outage window. The industry challenge is not eliminating emergent work but managing it with speed, evidence, and confidence (International Atomic Energy Agency, 2002; Electric Power Research Institute, 2003).

Two analytical gaps motivate this paper: the first one focuses on pre-outage planning, which today lacks data-driven mechanisms able to predict which components are likely to generate unplanned work. In this respect, resources (e.g., parts, specialized crew, contingency schedule float) are not strategically pre-positioned for unplanned work that is very likely to occur. Second, when an unexpected activity is discovered, the outage manager must promptly decide whether to start the work now, defer it within the outage window, defer it post-outage, or escalate. This decision is made without a structured assessment of likely duration, schedule impact, regulatory constraints, and precedent from prior outages (International Atomic Energy Agency, 2006).

The methods described in this paper address both of these gaps with a data analytics architecture consisting of a proactive planning pipeline and a reactive execution pipeline that

share a common knowledge graph. From a Prognostics and Health Management (PHM) point of view, the distinctive contribution of this paper is the integration of three capabilities into a single operational workflow: prognostic leading-indicator detection, data-driven uncertainty quantification, and evidence-based outage decision support. The framework integrates three individual contributions: 1) a temporal trend analysis stage that operates as a prognostic indicator for emergent work generation: detecting components whose degradation trajectories are escalating toward first failure before any emergent work record exists, extending health-index monitoring to the zero-emergent-history regime; 2) an uncertainty quantification method designed to determine the distribution associated with the completion time of an activity based on semantic analog retrieval and relevance-weighted percentile fitting that produces remaining-work-duration estimates at the p_{50} and p_{80} levels as a direct input to schedule risk management; and 3) a cross-phase impact model demonstrating that planning-phase pre-staging directly reduces execution-phase critical-path exposure. The combination of these three layers (prognostics feeding uncertainty quantification feeding decision support) is described in this paper. Prior works have addressed these layers individually: degradation trend monitoring and remaining-useful-life estimation (Jardine, Lin, & Banjevic, 2006; Si, Wang, Hu, & Zhou, 2011) provide the foundation for the planning-phase leading-indicator component, and the analog retrieval component builds on data-driven condition-based maintenance methods (W. Wang, Hussin, & Jefferis, 2012). The contribution here is their integration into a unified outage decision-support workflow with explicit cross-phase coordination.

The scheduling engine responsible for evaluating the schedule impact assessment (resource-constrained project scheduling, critical-chain buffers, and Monte Carlo propagation) is described separately in (Mandelli, 2026). This paper focuses on the analytics layers: prognostic trend detection, analog retrieval, uncertainty quantification, and evidence-based recommendation synthesis.

2. FRAMEWORK ARCHITECTURE

The framework consists of two pipelines operating on a shared knowledge graph as shown in Fig. 1. The planning pipeline runs in the months before the outage window opens. It processes historical condition reports (CRs) and work orders (WOs) from historic outage data (through natural language processing methods [NLP] (Sawicki, Ganzha, & Paprzycki, 2023; C. Wang, Mandelli, & Cogliati, 2024)), it constructs a component knowledge graph, and it produces a ranked risk register of components likely to generate emergent work.

The execution pipeline runs during the outage itself. Given a CR generated during the outage, it produces a structured recommendation (PROCEED, DEFER, ESCALATE, or MONITOR) with a quantified schedule impact, regulatory constraint assessment, and confidence-tiered duration estimate.

Note that the two pipelines share the same knowledge graph, so evidence accumulated during the planning phase (composite risk scores, trend labels, recommended pre-staging actions) is directly available when the execution pipeline processes a CR for a previously flagged component.

3. PLANNING PIPELINE

3.1. Data Ingestion and NLP Extraction (Stages A and B)

In this section, the term *outage cycle* refers to one complete refueling outage; a *training cycle* is any historical outage cycle whose records are used to build the knowledge graph and calibrate the composite risk scores, as opposed to the holdout cycle reserved for evaluation.

Stage A loads and validates component lists, CRs, WOs, and outage schedule data. Each historical outage activity is tagged with an emergence category: DISCOVERY (unplanned find), SCOPE_EXPANSION, REGULATORY_DRIVEN, or SCHEDULE_OPTIMIZATION. This category is stored in the knowledge graph and it propagates to the execution pipeline (where it feeds Stage F option scoring). A REGULATORY_DRIVEN activity triggers a mandatory precheck that may block deferral options before any scheduling analysis runs. Regulated components are flagged at ingest for downstream constraint checking.

Stage B addresses the presence of plant-specific abbreviations using a nuclear abbreviation dictionary; it then applies a hybrid named entity recognition (NER) pipeline (C. Wang et al., 2024) designed to extract equipment identifiers, failure mode keywords, and CR/WO cross references. Lastly, it computes an unknown-token rate as a data-quality signal. Components with an unknown-token rate exceeding 0.25 are flagged for analyst review rather than passed to automated scoring. The threshold is a design parameter; at 0.25, a description with one unresolved abbreviation every four is redirected (i.e., data quality flag), a level at which NER extraction quality degrades enough to produce unreliable component identification. Descriptions below the threshold are passed through with the unknown-token rate reported as a data-quality signal to downstream stages.

3.2. Knowledge Graph Construction (Stage C)

Stage C initializes a knowledge graph (Liu, Fu, Xu, Shi, & Ren, 2023) that links components and equipment to their associated CRs, WOs, and maintenance activities across all training outage cycles. The graph is the common data storage between the two pipelines (see Figure 1). The execution pipeline queries the same graph at runtime, so evidence accumulated during planning is immediately accessible.

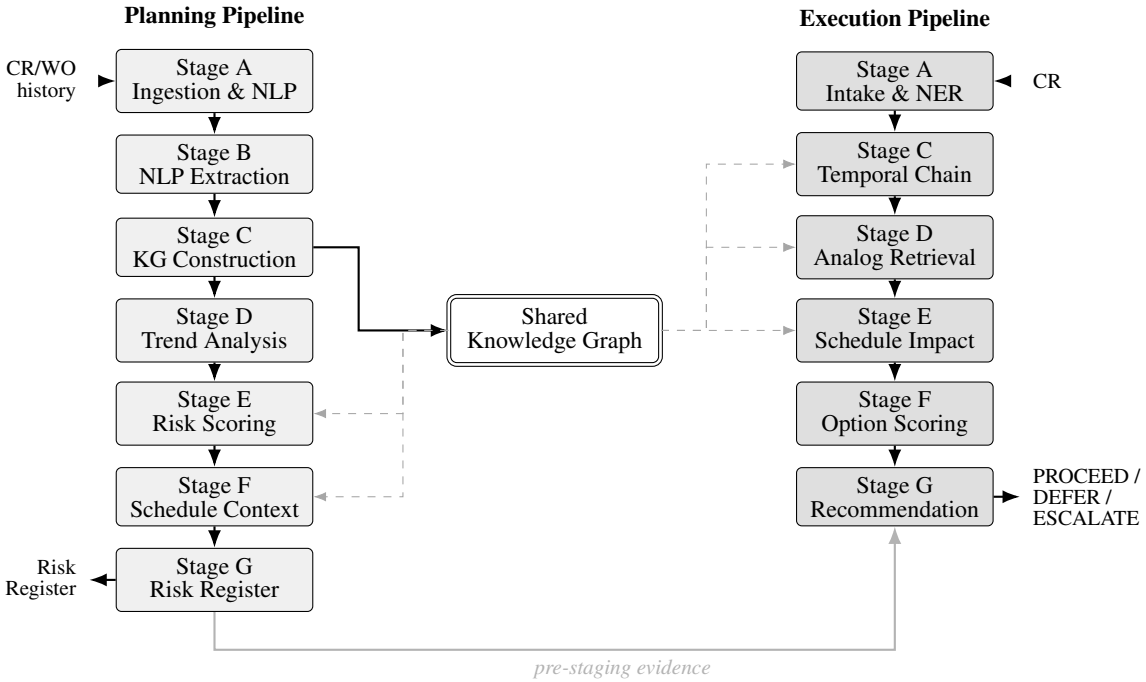

Figure 1. Cross-phase framework architecture. The planning pipeline (see pipeline on the left) processes historical data and outputs a risk register. The execution pipeline (see pipeline on the right) targets emergent activities. Both pipelines read from and write to a shared knowledge graph (center). The cross-phase arrows indicate that pre-staging evidence from the planning pipeline informs the execution-phase recommendation for flagged components.

Each component node stores its regulatory classification, criticality weight (derived from the probabilistic risk assessment importance ranking), system affiliation, and the sequence of CR and WO events across prior cycles.

3.3. Temporal Trend Analysis (Stage D)

Stage D is designed to address this question: is this component's degradation trajectory worsening across cycles, even before any emergent work has been recorded? This is the regime where evidence-based scoring (Stage E) would assign zero risk, because no emergent history exists to condition on. Stage D provides the early-warning signal that covers this gap. For each component, it analyzes three degradation signals across the N training outage cycles:

1. *Degradation CR frequency trend* $f_d(j)$: positive when the count of degradation-category CRs for component j is higher in the most recent training half than in the earlier half, capturing an increasing rate of degradation reports over time.
2. *Severity escalation* $e(j)$: positive when at least one training cycle records a CR severity escalation from observation to degradation relative to the preceding cycle for component j.
3. *Duration overrun trend* $r_d(j)$: positive when the mean overrun ratio across WOs for component j exceeds 1.15 *and* the most recent overrun ratio is higher than the earliest, confirming a worsening rather than a stable elevated baseline.

The composite trend score $T_j \in [0, 1]$ combines the three signals as a weighted indicator sum:

$$T_j = 0.40\mathbb{1}[f_d(j) > 0] + 0.30\mathbb{1}[e(j) > 0] + 0.30\mathbb{1}[r_d(j) > 0] \quad (1)$$

where $\mathbb{1}[\cdot]$ is the indicator function. The weights (0.40, 0.30, 0.30) and the escalation threshold were selected through structured expert elicitation; the same elicitation basis applies to the similarity weights in Eq. 4, the confidence-tier thresholds, and the decay constant λ. All are intended as starting-point values subject to site-specific recalibration rather than universally optimized parameters. A component is labeled *escalating* when $T_j \geq 0.5$, which requires at least two signals to be simultaneously non-zero. The key differentiator is that a component with *zero* prior emergent work records but with $f_d(j) > 0$ and $e(j) > 0$ receives $T_j = 0.70$ and is correctly flagged before any emergent activity exists. Without Stage D, such a component would be assigned zero risk under conventional evidence-based methods because no emergent record exists to condition it on.

3.4. Risk Scoring and Risk Register (Stages E–G)

Stages E–G synthesize the outputs of the preceding stages into actionable planning recommendations. The central question is: which components should be pre-staged before the outage begins, and at what confidence level? The answer depends on two complementary signals. The first is historical evidence: how often has this component exhibited degradation, and how reliably has degradation preceded emergent work in past cycles? The second is trend: is this component's degradation trajectory currently worsening, even if no emergent work has yet been recorded? First-failure components are not invisible to the planner because either path can trigger a risk register entry, as described below. Stage F assigns a confidence tier (DATA-SUPPORTED, SME-INFORMED, or MONITOR) and a recommended action; Stage G writes the resulting entries to the risk register that is shared with the execution pipeline.

Stage E computes a composite risk score from the historical record. Let D_j denote the event that component j had at least one degradation-category CR in a given training cycle, and let U_j denote the event that at least one emergent (unplanned) activity was recorded for component j in that cycle. The composite risk score is:

$$\mathrm{rs}_j = \frac{|\{c : D_j \in c\}|}{N} \times \frac{|\{c : U_j \in c,\ D_j \in c\}|}{|\{c : D_j \in c\}|} \times w_c(j) \tag{2}$$

where N is the number of training cycles, c indexes cycles, and $w_c(j) \in [0.5, 2.0]$ is a criticality weight obtained by linearly normalizing the component's risk importance measure from the plant probabilistic risk assessment model to $[0.5, 2.0]$, assigning 2.0 to the highest-importance component and 0.5 to the lowest.

A component receives a high score only when degradation is frequent, degradation reliably precedes emergent work, *and* the component carries high safety importance (a single weak factor suppresses the score, consistent with the conservative intent of a safety-critical risk register). When no training cycle records a degradation CR for component j ($|\{c : D_j \in c\}| = 0$), the second factor is undefined and rs_j is set to 0; the component is excluded from the risk register without raising an error.

Components with $\mathrm{rs}_j \geq 1.5$ are classified as DATA-SUPPORTED; those with $\mathrm{rs}_j \geq 0.5$ are SME-INFORMED. At maximum criticality ($w_c = 2.0$), these thresholds require the joint product of the first two factors to exceed 0.75 and 0.25, respectively; both values are design parameters calibrated on the training cycles and should be re-evaluated when the framework is applied to a dataset with a substantially different historical emergent work base rate. On the RF-22 training data, the two correctly flagged components sit at 0.2 and 0.3 score units inside their respective tier boundaries (RCP-A at 1.7 vs. the 1.5 threshold; SBR-14 at 0.8 vs. the 0.5 threshold), indicating that the tier assignments are stable under small perturbations of the training scores.

Table 1. Risk register tier assignment for the three RF-22 flagged components. Path codes: E = evidence ($\mathrm{rs}_j \geq$ threshold); T = trend ($T_j \geq 0.5$). †Escalation flag added; no emergent precedent.

Comp.	T_j signals	T_j	rs_j	Path	Tier
RCP-A	f_d, e, r_d	1.00	1.70	E + T	DATA-SUPPORTED
SBR-14	f_d	0.40	0.80	E	SME-INFORMED
CHG-03	f_d, e	0.70	0.00	T†	SME-INFORMED

Note that the trend score T_j (Eq. 1) and the composite risk score rs_j (Eq. 2) are *complementary*, not competing, mechanisms: rs_j is an evidence-based score that requires at least one prior emergent work record to be non-zero, while T_j is a trend-based signal that can flag a deteriorating component even before any emergent record exists. A component enters the risk register through either path: $\mathrm{rs}_j \geq 0.5$ (evidence-supported) or $T_j \geq 0.5$ (trend-escalating). When both conditions are met, the tier is determined by rs_j and the trend flag is reported as supplementary evidence. When only $T_j \geq 0.5$ is satisfied (no prior emergent work), the component is placed at SME-INFORMED tier with an escalation flag and an explicit note that no emergent precedent exists (requiring domain-expert confirmation before pre-staging resources are committed). This two-path structure is the mechanism that enables first-failure detection: a first-failure component has $\mathrm{rs}_j = 0$ by construction (no emergent record to populate the second factor of Eq. 2) but may reach $T_j \geq 0.5$ through degradation CR frequency growth and severity escalation. Table 1 illustrates the mechanism for the three components flagged in the RF-22 evaluation.

Stage F retrieves each flagged component's historical critical-path float consumption across prior cycles, quantifying how often the component has consumed critical-path time. Stage G assembles the ranked risk register with one finding and one recommended pre-staging action per component, delivered to the outage planning team weeks before the outage window opens.

4. EXECUTION PIPELINE

4.1. Intake and Entity Extraction (Stage A)

The execution pipeline starts from a CR description; from here, Stage A normalizes the CR text, expands abbreviations, and applies the same hybrid NER pipeline used in the planning phase. The output includes the component and system identifiers, a classified emergence type, detected regulatory

constraint keywords (technical specification identifiers, regulatory commitment references, corrective action program tags, hold points, and as low as reasonably achievable constraints), and a data-quality score.

Components with regulatory constraint keywords trigger a precheck that evaluates whether any defer option is admissible under technical specification surveillance requirements before any scheduling analysis proceeds.

4.2. Temporal Event Chain Scoring (Stage C)

When a new emergent CR is discovered during outage execution, a key issue to address is whether the failure is causally connected to prior events on the same component. A failure with a clear causal predecessor (e.g., a degradation CR that was active or recently closed when the emergent activity was discovered) provides stronger evidence for the analog retrieval and for the confidence assigned to the recommended action than an isolated, first-occurrence event. Stage C assesses this causal posture by retrieving the component's event history from the knowledge graph and scoring the temporal relationship between each prior event and the current emergent activity.

Allen interval algebra (Allen, 1983) provides the vocabulary for this classification: it defines seven exhaustive, mutually exclusive relations between two time intervals (PRECEDES, OVERLAPS, CONTAINS, SIMULTANEOUS, DURING, FOLLOWS, UNKNOWN), each of which carries a different implication about whether the prior event could plausibly have contributed to the current failure. The emergent activity is represented as an interval $[t_s, t_s + d]$ where t_s is the discovery timestamp and d is the planned duration supplied with the intake CR at Stage A; if no planned duration is available, the interval degrades gracefully to a point event ($d = 0$), and the Allen algebra proceeds on the discovery timestamp alone, with no dependency on the duration estimate produced later by Stage D. Each prior event is represented by its own interval $[t_i, t_i + d_i]$.

The seven Allen relations are mapped to causal relevance scores reflecting the strength of temporal coupling expected in maintenance sequences (Table 2).

A link's causal strength is the product of its relation score and a data-quality confidence factor:

$$\mathrm{cs\_link}(i) \;=\; r(\rho_{i,j}) \;\times\; \left(0.55\, q_i + 0.45\, e^{-\lambda\, \Delta t_i}\right) \qquad (3)$$

where $r(\rho_{i,j})$ is the Allen relation score, $q_i \in [0, 1]$ is the data-quality score of the prior event, Δt_i is the onset lag in hours from the end of the prior event to the start of the emergent activity, and λ is a decay constant set to $0.01\,\mathrm{h}^{-1}$, corresponding to a causal relevance half-life of approximately 69 h. This was chosen to match the timescale of within-outage maintenance sequences; a prior event 3 days before the emergent activity retains 50% causal weight, while events from prior outage cycles (months earlier) decay to negligible contribution. Links with $\mathrm{cs\_link} \geq 0.75$ are labeled *strong*; those ≥ 0.40 are *moderate*.

Table 2. Allen relation scores used in Stage C causal strength computation.

Relation	Score	Rationale
OVERLAPS	0.90	Prior event still active at discovery; strongest shared-mechanism evidence
CONTAINS	0.85	Prior event encloses emergent activity; broad maintenance context
PRECEDES	0.80	Classical prior-cause: earlier event set conditions for current failure
SIMULTANEOUS	0.50	Ambiguous causal direction between concurrent findings
DURING	0.30	Prior event more likely a concurrent discovery
FOLLOWS	0.10	Prior event post-dates emergent activity
UNKNOWN	0.00	No temporal information; contributes no evidence

Stage C produces a causal posture that propagates to the final recommendation confidence tier: *established* (at least one strong link), *partial* (at least one moderate link), or *insufficient_data* (no moderate links). In the three evaluated scenarios, causal posture does not serve as a decision gate but modulates the confidence tier of the recommendation. Its role as a direct decision input would surface in scenarios where a temporally linked prior event is still active at discovery, making the causal chain a factor in the feasibility of deferral options; this use case is structural in the current design and a subject of ongoing validation.

4.3. Historical Analog Retrieval and Uncertainty Quantification (Stage D)

Once an emergent CR has been normalized and its causal context assessed, the execution pipeline answers this question: how long will the repair take? Duration uncertainty is the primary driver of schedule risk; an activity that runs 30% over its estimated duration on a critical path can cascade into a critical outage extension. Direct estimation from first principles is infeasible without a component-level physics model, so Stage D takes an evidence-based approach: it retrieves historical WOs whose activity descriptions, component families, and task contexts are most similar to the current emergent CR, and derives calibrated duration estimates from the distribution of their actual durations.

The output is a pair of weighted percentile estimates (p_{50} and p_{80}) and a confidence tier determined by analog coverage. These estimates are passed to Stage E, which uses them as

inputs to a Monte Carlo schedule simulation to quantify the emergent activity's impact on the outage critical path.

Given the normalized activity description, Stage D retrieves analogous historical WOs using a multi-signal similarity score:

$$S(q,h) = 0.20\, s_{\text{lex}}(q,h) + 0.40\, s_{\text{sem}}(q,h) + 0.40\, s_{\text{ctx}}(q,h) \tag{4}$$

where s_{lex} is a lexical similarity score (BM25 variant (Robertson & Zaragoza, 2009)), s_{sem} is a dense embedding cosine similarity over sentence-level representations (Reimers & Gurevych, 2019), and s_{ctx} incorporates component family, task family, system, and discipline. Semantic and contextual components are weighted equally (0.40 each) because CR vocabulary varies widely for similar activities while component family is a strong duration discriminator; lexical overlap is largely redundant with the semantic score when embeddings are available. All weights are design parameters calibrated on Millbrook training cycles. Candidates below a similarity threshold of 0.35 are discarded.

Retrieved analogs are subject to outlier removal before distribution fitting. Each retained analog carries a relevance weight $w_i = S(q, h_i)$. The duration distribution is fitted as a weighted empirical distribution using the Type-7 interpolated quantile method:

$$Q_p \;=\; \text{WPercentile}_{w_i}(\{d_i\},\, p) \tag{5}$$

where d_i is the actual duration of analog i and $p \in \{0.50, 0.80\}$. The weighted Type-7 method normalizes the relevance weights w_i to sum to one, assigns each analog a weighted cumulative probability, and applies standard Type-7 linear interpolation between the two analogs whose cumulative weights bracket the target quantile p. High-similarity neighbors receive proportionally more weight, so the p_{50} and p_{80} estimates are dominated by the closest analogs rather than treating all retrieved matches as equally informative.

The confidence tier is assigned from the analog count n_a:

$$\text{tier} \;=\; \begin{cases} \text{DATA-SUPPORTED} & n_a \geq 5 \\ \text{SME-INFORMED} & 1 \leq n_a < 5 \\ \text{LOW-CONFIDENCE} & n_a = 0 \end{cases} \tag{6}$$

When $n_a = 0$, no duration estimate is produced and the final recommendation defaults to MONITOR, preventing fabricated precision in novel failure scenarios. The p_{80} estimate (not the p_{50}) is used to anchor contingency buffers, consistent with schedule risk guidance in project management practice.

Fig. 2 shows the fitted distribution for the RCP seal scenario (five analogs, DATA-SUPPORTED), illustrating how the weighted-percentile fitting produces calibrated p_{50} and p_{80} estimates from the retrieved analog pool.

4.4. Schedule Impact Assessment (Stage E)

Assessing how long a repair will take is necessary but not sufficient: the scheduler also needs to know whether inserting that repair will threaten the outage end date, and by how much. An activity that consumes all available float on a near-critical path can become a critical-path driver even if its duration estimate is accurate. Stage E loads the live outage schedule network, inserts the emergent activity at the proposed start point, and propagates the duration uncertainty from Stage D through a Monte Carlo simulation (1000 samples by default). A lognormal distribution is fitted to Stage D's p_{50} and p_{80} estimates (using $\Phi^{-1}(0.80) \approx 0.842$ to recover the lognormal parameters); sampled durations are propagated through the resource-constrained schedule network, with fixed permit lead times (radiation protection survey, scaffold erection, and electrical clearance) added before each propagation. The fraction of samples in which the emergent activity carries zero float is reported as the critical-path sensitivity score. Output metrics (p_{50}/p_{80} project finish times, drag, and sensitivity score) are passed to Stage F to score and rank the available insertion options; an activity with positive drag leaves Stage F with few feasible deferral choices, while ample float permits safe deferral or scope reduction. The resource-constrained critical-path engine is described in (Mandelli, 2026); Stage E treats it as a callable service.

4.5. Option Scoring and Recommendation Synthesis (Stages F and G)

With duration uncertainty quantified and schedule impact assessed, the final question is: what should the crew do right now? This is not simply a matter of picking the lowest-risk action (i.e., the feasible option set is itself constrained). Technical specification defer-prohibition rules, downstream resource conflicts, and regulatory hold-point requirements can eliminate one or more options before any scoring takes place, leaving the scheduler with a reduced menu of permissible choices. Stage F resolves this by first pruning infeasible options and then scoring the remaining ones; Stage G selects the lowest-risk feasible option and packages it into a recommendation structured for human review.

A key design requirement is traceability: in a nuclear setting, a recommendation without an auditable evidence chain is not actionable. Every output therefore includes the supporting historical records, analog count, confidence tier, and analyst attention flags (so that the operator can accept, override, or escalate the recommendation with full visibility into its basis).

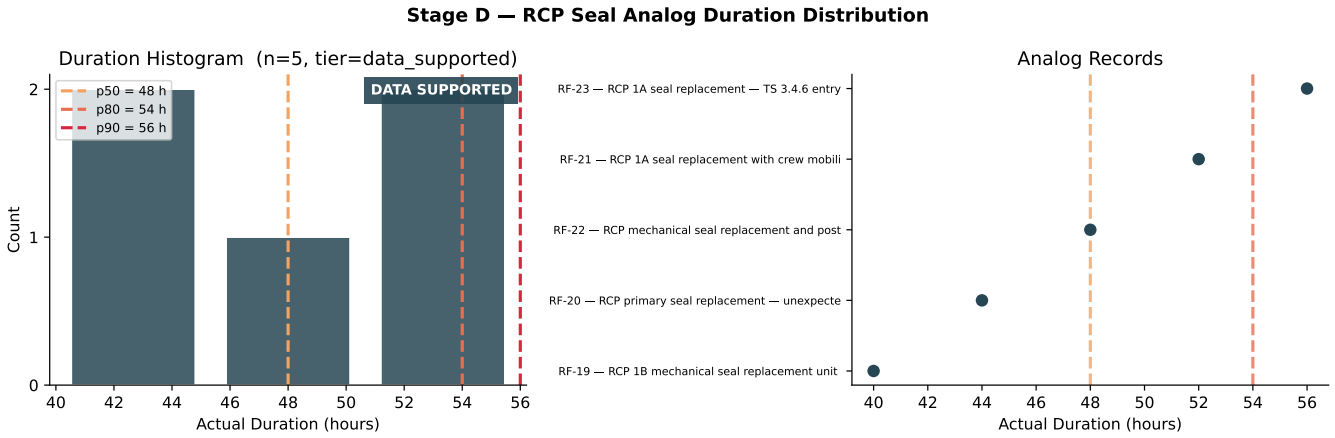

Figure 2. Stage D analog duration distribution for the RCP Train-A seal leak scenario. Five analogs are retrieved at DATA-SUPPORTED tier, producing $p_{50} = 48$ h and $p_{80} = 54$ h ($p_{90} = 56$ h shown for context). Each analog carries a relevance weight proportional to its multisignal similarity score (Eq. 4); higher-weight analogs dominate the percentile estimates. The p_{80} value is used as the contingency anchor for schedule impact assessment in Stage E. Analog labels in the right panel are abbreviated to fit.

Stage F generates all viable insertion options (insert now, defer in-outage, defer post-outage, scope reduction, and escalate) and scores each option's risk on $[0, 1]$. Options that violate technical specification defer-prohibition constraints are blocked before scoring. Options using a defer action are further checked for downstream resource conflicts and regulatory hold-point conflicts.

Stage G selects the lowest-risk feasible option and maps the decision to a status label: PROCEED, DEFER, ESCALATE, or MONITOR. The recommendation is accompanied by (a) an executive summary with plain-language rationale; (b) an evidence chain linking each claim to its source CR, WO, or schedule artifact; (c) a confidence tier that reflects both the analog retrieval quality (Stage D) and the causal posture (Stage C); and (d) analyst attention flags that mandate human review when regulatory constraints are present, when the confidence tier is LOW-CONFIDENCE, or when upstream data quality is below the threshold.

5. CROSS-PHASE INTEGRATION

The planning and execution pipelines are designed to function independently: outage planning runs months before the outage opens on historical data alone, and execution runs in real time as emergent CRs arrive. The integration between them is not a data handoff at a fixed point in time but a shared knowledge graph that is continuously readable by both pipelines. When execution processes an emergent CR, the planning-phase risk register entries for that component (i.e., composite risk score, trend label, confidence tier, and a record of whether the recommended pre-staging action was taken) are already present in the graph and are incorporated into the Stage G evidence chain without any additional data transfer step.

If the planning pipeline flagged a component and the pre-staging recommendation was acted on, crew and parts are already on site when the emergent CR arrives. This eliminates the mobilization and parts-procurement delays that would otherwise be incurred before physical work can begin, directly reducing the critical-path extension caused by the emergent activity.

The impact delta for a given component is:

$$\Delta T_{\mathrm{CP}} \;=\; T_{\mathrm{mob}} + T_{\mathrm{parts\text{-}wait}} \tag{7}$$

where T_{mob} is the crew mobilization time and $T_{\mathrm{parts\text{-}wait}}$ is the parts procurement delay, both of which are reduced to zero when pre-staging actions have been executed. This quantity is computed at the recommendation time and reported alongside the schedule impact estimate to provide a measurable return-on-investment signal for the planning-phase analytics investment.

6. EVALUATION

The evaluation tests three hypotheses: first, that the execution pipeline produces the correct decision for scenarios representing each distinct decision branch (regulatory block, float-based deferral, and zero-analog abstention). Second, that the planning pipeline correctly identifies, before the outage opens, all components that will generate emergent work in the holdout cycle, including components with no prior emergent work record. Third, that acting on a planning-phase flag

reduces the execution-phase critical-path extension relative to a fully reactive baseline.

6.1. Dataset and Experimental Setup

The framework is evaluated on the Millbrook synthetic nuclear plant dataset, a purpose-built dataset of 150 WOs across six simulated outage cycles with a 22-component equipment list covering four systems. The dataset was constructed by generating WO and CR records with failure-mode patterns representative of three degradation categories: (i) components with escalating degradation CRs and prior emergent work history (DATA-SUPPORTED tier candidates), (ii) components with a single prior emergent event and moderate degradation trend (SME-INFORMED tier candidates), and (iii) components with escalating degradation trends but no prior emergent work record (first-failure candidates, for exercising the Stage D path). CR text descriptions were generated to test the NLP extraction pipeline across a range of abbreviation density and technical vocabulary.

6.2. Execution Pipeline

Three scenarios are evaluated, each designed to exercise a distinct branch of the developed logic. The pipeline receives a CR as input; no decision label is provided.

Scenario 1: RCP Train-A seal leak (regulatory block path). The CR reads: *"RCP Train-A mechanical seal leaking approximately 2 GPM. TS 3.4.6 entry required. Hold point placed on WO-44821 pending operability determination. Seal replacement estimated 48 h."* The component (Reactor Coolant Pump Train A, Reactor Coolant System) has three prior degradation CRs across RF-21 through RF-23 documenting progressively worsening seal leakoff flow. Five analogs of prior RCP mechanical seal replacements are available in the knowledge base. The expected decision is ESCALATE: the technical specification identifier TS 3.4.6 prohibits deferral, and the activity lies on the critical path with zero remaining float.

Scenario 2: Snubber scope expansion (float-based deferral path). The CR reads: *"While performing scheduled snubber inspection on line RCS-SN-204, technician identified two additional snubbers requiring functional testing. Opportunistic scope addition to existing WO-44610. Additional work estimated 8 h."* The component (snubber RCS-SN-204) has one prior CR and one prior PM; no regulatory keywords appear in the CR text. Five analogs of snubber functional tests are available. The expected decision is DEFER: the activity is non-safety-critical with 28 h of schedule float remaining after insertion and no critical-path impact.

Scenario 3: Unknown drain valve leak (zero-analog abstention path). The CR reads: *"During routine equipment walkdown, drain valve on auxiliary feedwater system header (AFWS-DRN-033) found with active packing leak. No prior condition reports or work orders found for this valve assembly. Component not previously catalogued in maintenance history."* The component has no prior CRs, no prior WOs, and no knowledge graph entry. Zero analogs are retrieved. The expected decision is MONITOR: without an analog pool, the pipeline cannot produce a calibrated duration estimate and the zero-analog policy prevents a fabricated recommendation.

Table 3. Execution pipeline evaluation on three holdout scenarios.

Scenario	Expected	Actual	Analogs	p_{80} (h)
RCP Train-A seal leak	ESCALATE	ESCALATE	5	54.0
Snubber scope expansion	DEFER	DEFER	5	9.5
Unknown component leak	MONITOR	MONITOR	0	—

Expected and actual decisions are compared in Table 3.

All three scenarios produce the expected decision, each exercising a distinct branch of the triage logic: regulatory block, float-based deferral, and zero-analog abstention. These results constitute a scenario-based demonstration that each decision branch executes correctly on representative inputs; they are not a statistical accuracy benchmark over a large holdout set. Fig. 3 summarizes the three scenarios across all primary pipeline metrics.

6.3. Planning Pipeline

The planning pipeline evaluation tests the following hypothesis: does the pipeline correctly flag, before the outage opens, all components that will generate emergent work in the holdout cycle (including components with no prior emergent work record)?

The pipeline is trained on five outage cycles (RF-17 through RF-21) and evaluated on the RF-22 holdout. Three components are of interest in this evaluation:

- **RCP-A**: has two prior emergent work records and all three trend signals active ($T_j = 1.0$); a strong evidence-path candidate.
- **SBR-14**: has one prior emergent record and a moderate degradation trend; a weaker evidence-path candidate.
- **CHG-03**: has an escalating degradation CR frequency and severity trend but *no prior emergent work record* ($\mathrm{rs}_j = 0$ by construction); the only way it can enter the risk register is via the trend path ($T_j \geq 0.5$).

RCP-A and SBR-14 both generate emergent work in RF-22; CHG-03 does not. A successful result requires all three to appear in the risk register (RCP-A and SBR-14 via the evidence path, and CHG-03 via the trend path) demonstrating that the

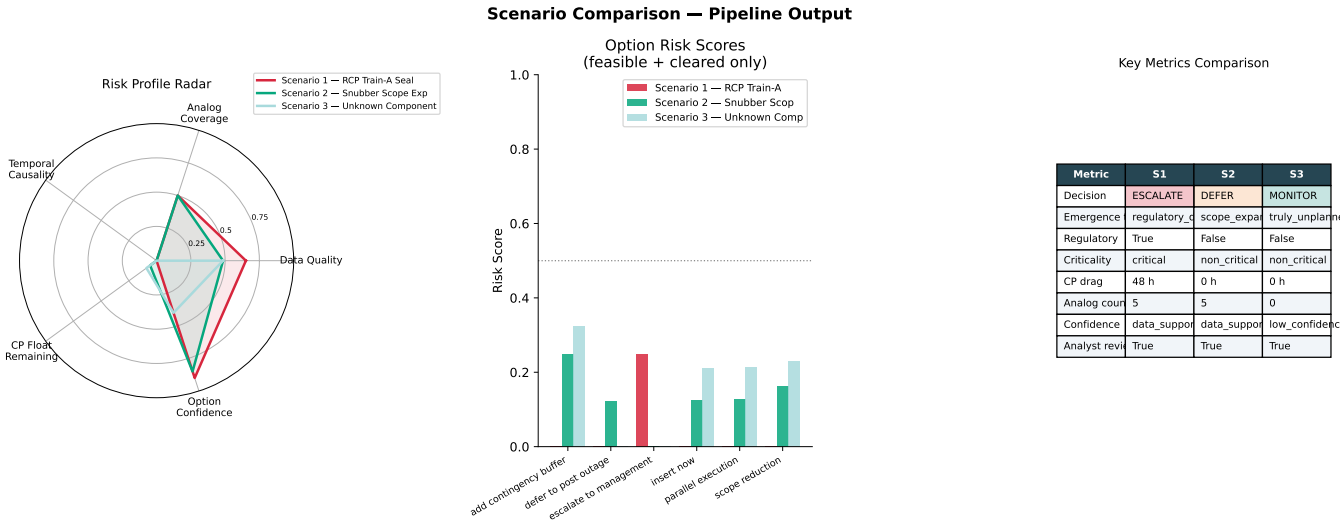

Figure 3. Cross-scenario comparison across five pipeline metrics for the three evaluation cases: critical-path drag, analog count, confidence tier, number of feasible insertion options, and final decision status. The contrast between the ESCALATE case (high CP drag, regulatory block), the DEFER case (zero CP drag, data-supported tier), and the MONITOR case (zero analogs, no estimate) illustrates how the pipeline adapts its reasoning path to the input evidence rather than applying a fixed decision rule.

pipeline's coverage extends to the first-failure regime. The ranked risk register output is shown in Fig. 4.

Both components that generated emergent work in RF-22 appear in the risk register: RCP-A at DATA-SUPPORTED tier (composite risk score 1.7, two prior emergent records plus all three trend signals active) and SBR-14 at SME-INFORMED tier (composite risk score 0.8, one prior emergent record). CHG-03 is flagged at SME-INFORMED tier via the trend path ($T_j = 0.70$, $\mathrm{rs}_j = 0$) without generating emergent work in RF-22; the Stage D mechanism fires correctly in the sense that a genuine escalating degradation trend exists, but this is not a validated first-failure prediction; it demonstrates that the trend-path entry condition activates when intended. On this single holdout cycle the Stage D mechanism fires for both emergent-work components (2 of 2; statistical significance not assessable at this sample size) and one additional component with an escalating trend; these results characterize mechanism behavior on a 22-component synthetic dataset, not predictive performance, and generalization is addressed in the Discussion.

6.4. Cross-Phase Impact

The third hypothesis tests whether acting on a planning-phase flag produces a measurable reduction in critical-path extension relative to a fully reactive baseline. The comparison requires a component that (a) was correctly flagged by the planning pipeline, (b) generated emergent work in the holdout cycle, and (c) would have incurred a mobilization delay under reactive operations. RCP-A satisfies all three conditions and is used as the anchor component. SBR-14 is excluded because its emergent activity was correctly deferred with 28 h of remaining schedule float, making the cross-phase impact zero by construction.

Two scenarios are compared:

- **Reactive baseline**: no planning-phase flag is acted on; parts and crew are not pre-staged; the full mobilization delay ($T_{\mathrm{mob}} = 8.5$ h, drawn from the Millbrook WO history for RCP-class repairs) must be incurred before physical work can begin.
- **Pre-staged**: the planning-phase flag is acted on weeks before the outage; parts and crew are on site at discovery; mobilization delay is zero.

The critical-path extension in each case is the sum of mobilization delay and repair activity duration. The cross-phase saving is the reduction in that extension attributable solely to eliminating the mobilization delay.

The 37% reduction in critical-path extension is entirely attributable to eliminating the assumed 8.5 h crew mobilization delay (a scenario parameter representing a realistic but conservative estimate for specialized nuclear maintenance mobilization) for a component that had been identified weeks earlier by the planning pipeline. The 14.5 h residual reflects the irreducible execution time of the repair work itself, which pre-staging cannot compress. More generally, the savings fraction scales as $T_{\mathrm{mob}}/(T_{\mathrm{mob}} + T_{\mathrm{activity}})$; for the 14.5 h repair activity, mobilization delays ranging from 4 h to 16 h yield savings of 22% to 52%, giving practitioners a direct way to estimate the pre-staging benefit for their own site mobilization parameters.

Pre-Outage Risk Register — 6 Weeks Before RF-22

Rank	Component ID	Description	System	Tier	Trend	Reg ★	Score	Recommended Action
1	1RHS-P-001A	RHR Pump 1A	Residual Heat Remova	DATA-SUPPORTED	escalating	★	2.00	Pre-order parts; allocate contingency crew
2	1RHS-E-001A	RHR Heat Exchanger 1A	Residual Heat Remova	SME-INFORMED	escalating	★	2.00	Stage spare; SME walkdown on day 1
3	1CSP-P-001B	Containment Spray Pump 1B	Containment Spray	SME-INFORMED (T)	escalating	★	0.00	Stage spare; SME walkdown on day 1
4	1CCW-P-002A	Component Cooling Water Pump	Component Cooling Wa	—	insufficient		0.00	No action required
5	1RHS-V-001A	RHR Suction Isolation Valve	Residual Heat Remova	—	insufficient	★	0.00	No action required

Millbrook synthetic holdout evaluation

Figure 4. Planning pipeline risk register for the RF-22 holdout outage cycle (Millbrook synthetic dataset). Components are ranked by composite risk score (Eq. 2). Confidence tier is color-coded: dark teal = DATA-SUPPORTED and medium teal = SME-INFORMED. The escalation flag on the third-ranked component (no prior emergent record) is generated exclusively by the Stage D temporal trend score ($T_j \geq 0.5$), demonstrating the first-failure detection path described in Section 6.3.

7. DISCUSSION

The framework's most distinctive property is its response when Stage D retrieves no historical analogs for the incoming CR: rather than falling back to a population-level statistic, the recommendation defaults to MONITOR and no duration estimate is produced. This behavior is referred to here as the *zero-analog policy*. This design choice prioritizes calibration over coverage. An alternative (using population-level statistics as a fallback) would avoid the MONITOR outcome but would produce estimates with no traceability to the specific failure mode at hand. For safety-critical maintenance decisions, the absence of an estimate is more informative than a poorly grounded one.

The analog count is a direct confidence signal to the analyst. An SME-INFORMED recommendation with four analogs requires domain-expert validation before commitment; a DATA-SUPPORTED recommendation with 15 analogs can be acted on with higher confidence. This explicitness distinguishes the framework from approaches that report a single confidence value with no indication of its data basis. The zero-analog policy is simultaneously a strength and a coverage constraint: on a small or specialized WO index, a substantial fraction of novel failure modes may fall into the MONITOR regime. Quantifying this coverage fraction on a real plant dataset is a necessary step before operational deployment; it directly bounds the fraction of field CRs for which the framework can provide a duration estimate.

A limitation of the current evaluation is dataset scale. The Millbrook dataset contains 150 WOs across several simulated outage cycles; generalization claims require an evaluation on multiyear plant histories spanning hundreds of components, which is the focus of ongoing work with plant data partners. The cross-phase scenario (Section 6.4) grounds the planning pipeline's contribution in operational terms: the planning-phase flag produces a measured reduction in execution-phase critical-path impact, making the value of pre-staging directly observable.

A second potential concern is cross-plant adaptation. The framework depends on plant-specific inputs (an abbreviation dictionary, a component taxonomy, and technical specification constraint mappings) that must be configured for each site. Critically, this is a *knowledge-engineering* cost, not a retraining cost; the retrieval weights, Allen relation scores, confidence tier thresholds, and composite risk score formula are not learned from plant data and do not require re-estimation when the framework is deployed at a new site. The unknown-token rate (Stage A), data-quality score, and confidence tier propagation are designed precisely to surface cases where the plant-specific configuration is incomplete, routing those cases to MONITOR or analyst review rather than allowing a poorly grounded recommendation to reach the outage manager.

The decay parameter λ and the Allen relation scores in Stage C are the two judgment-sensitive inputs whose values are not validated by the current evaluation. The chosen $\lambda = 0.01\,\mathrm{h}^{-1}$ places the causal relevance half-life at 69 h; at an order of magnitude smaller ($0.001\,\mathrm{h}^{-1}$) the decay would be negligible over an outage window, while at an order of magnitude larger ($0.1\,\mathrm{h}^{-1}$, half-life 7 h) only events within a single shift would retain substantial weight. The 69 h value is consistent with within-outage maintenance cadences and can be adjusted for plants with substantially shorter or longer activity sequences without modifying any other model component.

The remaining design weights (i.e., trend score coefficients

(Eq. 1), similarity weights (Eq. 4), and Allen relation scores) are monotone in their qualitative direction: increasing the weight of any single signal raises the score of components that are positive on that signal and leaves all others unchanged. A systematic sensitivity study is deferred to the plant-scale evaluation. On the Millbrook dataset, perturbing each weight in Eq. 1 and Eq. 4 by $\pm 20\%$ leaves all three decisions unchanged: the RCP-A scenario is resolved by a regulatory block (categorical, independent of score magnitudes), the snubber scenario by a float comparison that remains positive under any plausible weight variation, and the unknown scenario by the zero-analog policy. The tier boundaries for the planning pipeline are similarly stable: the 0.2 and 0.3 unit margins between the scored components and their respective thresholds (Section 6.3) absorb small weight perturbations without changing tier assignments.

The framework currently treats the knowledge graph as an in-memory structure rebuilt at the start of each outage cycle. Integration with a live plant Computerized Maintenance Management System (where CRs and WOs are written continuously) would allow the execution pipeline to operate on up-to-date evidence without a cycle-boundary rebuild.

8. CONCLUSIONS

This paper presented a cross-phase framework that integrates prognostic leading-indicator detection, calibrated uncertainty quantification, and evidence-based decision support into a single operational workflow for emergent work management in nuclear refueling outages. The integration is the primary contribution: each layer informs the next, and the planning-phase prognostic output directly reduces execution-phase critical-path exposure in a way that neither layer could achieve alone.

Within that integration, Stage D provides a leading-indicator prognostic capability for emergent work generation: by detecting escalating degradation trajectories before any emergent record exists, it extends health-index monitoring to the zero-emergent-history regime and enables pre-staging of resources for high-probability first failures. Analog retrieval assigns calibrated remaining-work-duration estimates at the p_{50} and p_{80} levels via relevance-weighted percentile fitting, with an explicit analog count as a confidence signal. Allen interval algebra classifies causal posture between prior events and the current emergent activity and is structurally wired into the confidence propagation path; demonstrating its effect on recommendation outcomes is a direction for future evaluation. Cross-phase integration enables planning-phase pre-staging to directly reduce execution-phase critical-path exposure, demonstrated at 22–52% savings across the plausible mobilization-delay range (4–16 h), with 37% at the nominal 8.5 h value.

The framework is grounded in a zero-analog policy; when retrieval produces no comparable historical precedent, the system returns MONITOR rather than a fabricated estimate, preserving decision quality under novel failure conditions. Future work will focus on evaluation at plant scale, integration with live CMMS data streams, and extension of the temporal trend analysis to multi-cycle prognostic horizon estimation.

ACKNOWLEDGMENT

This work was supported by the U.S. Department of Energy, Office of Nuclear Energy, under the Light Water Reactor Sustainability (LWRS) program. AI tools assisted with the development of the presented framework and the revision of this paper. All technical analyzes, designs, results, and conclusions were developed and verified by the authors.

REFERENCES

Allen, J. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM*, *26*(11), 832–843.

Electric Power Research Institute. (2003). *Outage management benchmarking guideline* (Tech. Rep. No. EPRI 1004383). Palo Alto, CA: EPRI.

International Atomic Energy Agency. (2002). *Nuclear power plant outage optimisation strategy* (Tech. Rep. No. IAEA-TECDOC-1315). Vienna, Austria: IAEA.

International Atomic Energy Agency. (2006). *Indicators for management of planned outages in nuclear power plants* (Tech. Rep. No. IAEA-TECDOC-1490). Vienna, Austria: IAEA.

Jardine, A. K. S., Lin, D., & Banjevic, D. (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mechanical Systems and Signal Processing*, *20*(7), 1483–1510. doi: 10.1016/j.ymssp.2005.09.012

Liu, R., Fu, R., Xu, K., Shi, X., & Ren, X. (2023). A review of knowledge graph-based reasoning technology in the operation of power systems. *Applied Sciences*, *13*(7). Retrieved from https://www.mdpi.com/2076-3417/13/7/4357

Mandelli, D. (2026). Risk-informed resource-constrained scheduling for refueling outage management. In *Proceedings of the 18th international probabilistic safety assessment and analysis (psam-18).* Pittsburgh, CA.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing, EMNLP-IJCNLP 2019, hong kong, china, november 3-7, 2019* (pp. 3980–3990). Association for Computational Linguistics.

Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, *3*(4), 333–389.

Sawicki, J., Ganzha, M., & Paprzycki, M. (2023, 08). The state of the art of natural language processing—a systematic automated review of nlp literature using nlp techniques. *Data Intelligence*, *5*(3), 707-749. doi: 10.1162/dint_a_00213

Si, X.-S., Wang, W., Hu, C.-H., & Zhou, D.-H. (2011). Remaining useful life estimation – A review on the statistical data driven approaches. *European Journal of Operational Research*, *213*(1), 1–14. doi: 10.1016/j.ejor.2010.11.018

Wang, C., Mandelli, D., & Cogliati, J. (2024). Technical language processing of nuclear power plants equipment reliability data. *Energies*, *17*(7). Retrieved from https://www.mdpi.com/1996-1073/17/7/1785 doi: 10.3390/en17071785

Wang, W., Hussin, B., & Jefferis, T. (2012). A case study of condition-based maintenance modelling based upon the oil analysis data of marine diesel engines using stochastic filtering. *International Journal of Production Economics*, *136*(1), 84–92. doi: 10.1016/j.ijpe.2011.09.016

BIOGRAPHIES

Diego Mandelli is an R&D Scientist in the "Plant Optimization" Department at the Idaho National Laboratory (INL). His areas of expertise include risk, reliability, and system health management. His research focuses on the development of probabilistic methods based on machine learning, data mining and optimization algorithms. He is currently employing these methods to perform a state-of-the-art simulation-based safety assessment (also known as dynamic probabilistic risk assessment), system health management, and stochastic optimization. His developed methods range from data preprocessing, system modeling, system analysis, data mining and visualization, and decision-making. He holds a Ph.D. degree in nuclear engineering from The Ohio State University (2011).

Congjian Wang is a computational nuclear scientist at INL. He received his B.E. degree in engineering physics and his Ph.D. degree in nuclear engineering and carried out postdoctoral research focused on machine learning, advanced sensitivity and uncertainty quantification, verification, validation, and data assimilation. He has more than 10 years of experience in developing and implementing artificial intelligence algorithms including but not limited to physics-based reduced order modeling, surrogate modeling, data mining techniques, and deep neural networks. He is the main developer of the Risk Analysis Virtual ENvironment (RAVEN) framework and is mainly responsible for developing and implementing supervised and unsupervised learning algorithms, cross validations, and verification and validation metrics within the RAVEN framework.