

Stopping Criteria in Simulation: A Survey of Methods, Applications, Challenges, and Future Directions

ARTUR LIGIERI NUNES¹, GUILHERME MOREIRA¹, JOAO P. A. DANTAS², AND LUIZ ALBERTO VIEIRA DIAS¹

¹Instituto Tecnológico de Aeronáutica, São José dos Campos, São Paulo, 12288-900, Brazil.

²Instituto de Estudos Avançados, São José dos Campos, São Paulo, 12228-001, Brazil.

Corresponding author: Artur Ligieri Nunes (e-mail: nunesaln@ita.br).

This work was supported by the Article Processing Charge funded by the Coordination for the Improvement of Higher Education Personnel (CAPES)-Brazil (ROR identifier: 00x0ma614).

ABSTRACT Determining when accumulated simulation output is sufficient for an inference or decision is a recurring problem because premature termination can compromise statistical properties, whereas unnecessary continuation consumes computational and sampling resources. This survey examines data-dependent termination under uncertainty: operational rules that update evidence from sequential observations or batches and decide whether acquisition should stop or continue. A protocol-driven search of Scopus and Web of Science, covering 2015-2025 through four targeted strategies, identified 54 primary studies presenting 95 operational contributions, defined as distinct stopping procedures or operationally distinct variants. Reporting was informed by PRISMA 2020 and PRISMA-S. The contributions were organized into five families according to their terminal rationale: precision or uncertainty sufficiency, discrete-decision sufficiency, risk or loss sufficiency, operational-state sufficiency, and resource-aware acquisition and stopping. Among the 95 contributions, demonstrated evidence was documented for finite-sample behavior in 82, comparison with alternative rules in 66, robustness in 17, computational or scalability assessment in 47, and implementation availability in 7. A conditional practitioner map relates nine application constraints to the five families without prescribing a universal rule. The synthesis identifies two cross-family gaps concerning external evidence and changing data streams, together with a corpus-bounded application gap in defense constructive simulation.

INDEX TERMS Stopping criteria, data-dependent termination, sequential stopping, simulation output analysis, Monte Carlo simulation, sequential estimation.

I. INTRODUCTION

SIMULATION enables model-based experimentation when intervention in the actual system is infeasible, too costly, or operationally disruptive, and it also permits the analysis of proposed systems that do not yet exist [1]. Its established applications span manufacturing, business, and service processes [2], including manufacturing-system design and operation [3]. Contemporary examples relevant to stopping decisions include offshore-infrastructure wave analysis [4], high-dimensional Bayesian Markov chain Monte Carlo [5], machine-learning model training [6], and defense-oriented simulation for operational analysis and decision support [7], [8].

Once simulation output can be generated repeatedly, the analyst must determine what to estimate, classify, or select and when further acquisition is no longer justified. In Monte Carlo mean estimation, too few replications can leave the sample mean and downstream analyses unrepresentative,

whereas too many consume computation time that could be used for further model analysis [9]. More generally, additional replications may improve confidence or precision but incur costs in time, computing power, and money [10]. The stopping problem therefore concerns whether the information expected from further sampling justifies its setting-specific cost.

The required replication count or run length cannot always be determined reliably before execution. For terminating simulations, pilot information can support an approximate replication-count calculation; for non-terminating simulations, the run length required for a prescribed accuracy is particularly model-dependent and may be difficult to anticipate even within an order of magnitude [1], [11]. Data-dependent procedures address this uncertainty by reassessing, as information accumulates, whether sampling should continue.

This survey defines its common object as *data-dependent*

termination under uncertainty: operational rules that update evidence from sequential observations or batches and decide whether acquisition should stop or continue. The stopping time may affect statistical validity, computational or sampling cost, and the quality of the subsequent decision. These are potential consequences of stopping, not mandatory requirements that every rule must evaluate or optimize.

Data-dependent termination also requires inferential care. Output retained after satisfying a nontrivial criterion is conditional on that selection event, and its distribution can differ from the original data-generating distribution [12]. This does not mean that every stopping rule biases every statistic. In the finite-sample absolute-precision confidence-interval procedures examined by Singham, sequential stopping changed achieved coverage relative to its nominal level, while the direction and magnitude of the effect depended on the rule and its parameters [12]. Stopping procedures should therefore be evaluated for the inferential quantity and data setting in which they are implemented, rather than assumed to share the same bias, coverage, or error behavior.

A. THEORETICAL BACKGROUND

Monte Carlo and sequential inference provide distinct parts of the historical basis for simulation stopping. Metropolis and Ulam presented an early published account of Monte Carlo as a statistical, computer-suited approach to difficult mathematical-physics calculations [13]. Chow and Robbins later established the asymptotic theory of fixed-width sequential confidence intervals for a population mean with unknown variance under independent observations [14]. Their result was not itself a simulation output-analysis method, but it supplied a foundation for later precision-driven sequential procedures.

Simulation-specific methods subsequently addressed dependence, run length, and replication count. Fishman developed an iterative sample-size procedure that estimated the variance of correlated simulation output and acquired further observations until a prescribed precision criterion for the process mean was met [15]. Lavenberg and Sauer formulated sequential relative-width stopping rules for regenerative steady-state simulation and established asymptotic confidence-interval validity as the requested width decreased [16]. Moeller and Welch addressed a related but distinct problem: constructing confidence intervals for correlated, covariance-stationary simulation output by estimating the sample-mean variance from the zero-frequency spectrum [17]. Their method belongs to output analysis rather than to data-dependent stopping.

By the early 1980s, Law and Kelton had surveyed and empirically compared sequential and fixed-sample procedures for steady-state simulation [18], [19]. This distinction remains important. A fixed-sample or fixed-horizon procedure determines the replication count or run length before the resulting output is observed. A sequential procedure instead updates the relevant statistic and continues acquisition until a pre-specified criterion is satisfied. Such procedures can help

meet a precision target, but their achieved coverage and sampling requirements remain dependent on their assumptions, parameters, and finite-sample behavior [1], [10].

Two termination levels must be distinguished. In a terminating simulation, a natural event specified before execution determines when an individual replication or scenario ends. A non-terminating simulation lacks such a natural endpoint and commonly targets long-run system behavior [11]. This classification depends on the objective of the study rather than only on the physical system being modeled [1].

Scenario termination does not, by itself, determine whether the collected simulation output is sufficient. A terminating scenario may still require a separate decision about how many independent replications to execute. Conversely, a non-terminating simulation may require a decision about the length of a single run, its warm-up period, the number of independent runs, or a combination of these choices [1], [11]. The focal object of this survey is this second level: the data-dependent decision to continue or stop acquiring evidence, not the modeled event that ends an individual scenario.

The reviewed contributions arise primarily in stochastic settings, where the output used for inference or decision-making varies across observations, replications, or sampled states. Deterministic algorithms and deterministic components are not excluded by label alone; they enter the focal corpus only when they participate in the explicit, data-dependent stop/continue structure defined by the review protocol. The relevant boundary is therefore operational, not a general taxonomy of simulation modalities.

Verification, validation, and output analysis address different sources of uncertainty in a simulation study. Verification concerns whether the conceptual model has been correctly translated into an executable program, while validation concerns whether the model adequately represents the actual system for the intended use. Output analysis instead quantifies uncertainty in estimating the performance measures of the simulation model, including through choices of warm-up, run length, and independent replications [1].

A stopping rule does not replace model verification or validation. Conversely, completion of a universal V&V stage is not treated here as an operational property shared by all 95 contributions. The survey examines how those contributions update evidence and decide whether further acquisition is required. Model credibility and output sufficiency remain related but distinct questions.

Despite their heterogeneity, the reviewed contributions share an abstract operational cycle. A new observation, replication, batch, or adaptive acquisition updates an estimator, uncertainty measure, error proxy, loss, risk, ranking, structural state, or resource condition. A data-dependent predicate then determines whether acquisition continues or stops. Continuing returns the process to acquisition; stopping produces a terminal estimate, interval, classification, selection, structural summary, control action, or decision (Figure 1).

This shared cycle is an analytical abstraction, not a claim of methodological equivalence. Contributions may differ ma-

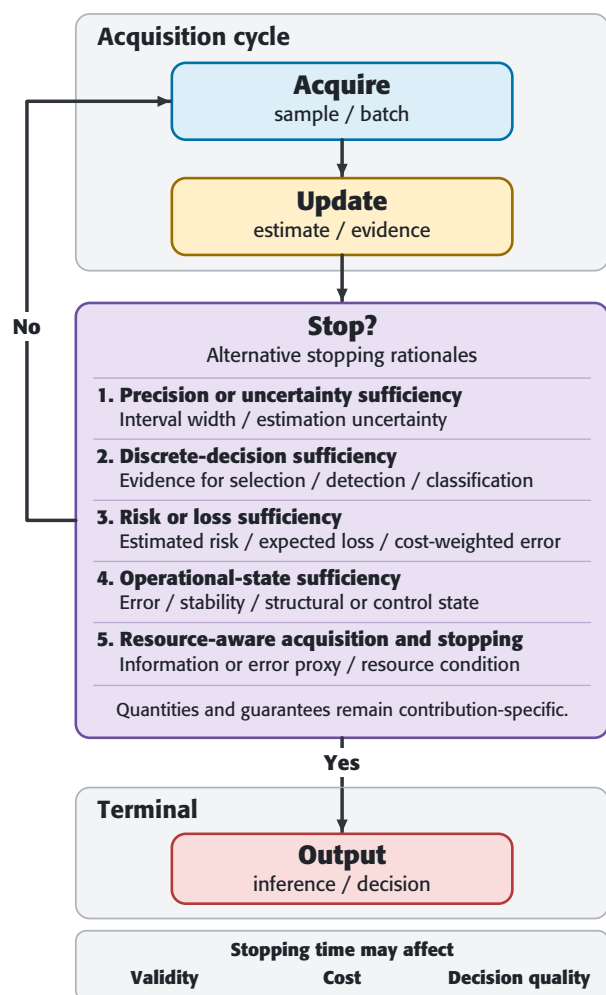


FIGURE 1. Common operational cycle for data-dependent termination under uncertainty. New information updates the monitored quantity before a stop/continue decision is evaluated. The five labels within the stop decision represent the five stopping-criterion families identified in this survey. Continuing returns the process to acquisition; stopping produces a terminal output. Validity, cost, and decision quality are potential consequences of the stopping time rather than universal inclusion requirements.

terially in their objective, monitored quantity, update schedule, terminal condition, output, assumptions, and guarantees. Neither family membership nor structural similarity transfers evidence or validity from one contribution to another.

B. OVERVIEW

Earlier work provides guidance on specific estimation problems. Law and Kelton [18], [19] survey and empirically compare sequential and fixed-sample procedures for confidence intervals for steady-state means. Bicher et al. [9] provide a focused technical review and numerical evaluation of replication-count methods for Monte Carlo mean estimation. These works address how to obtain a mean estimate with quantified uncertainty. The broader problem considered here also asks what justifies termination when the intended output is a classification, selection, operational state, or resource-

constrained result. This survey therefore compares stopping rationales and mechanisms alongside the evidence reported for their behavior and use. Table 1 summarizes this positioning.

This survey selected 54 primary studies and identified 95 operational contributions. A study was the unit of selection, whereas an operational contribution was the unit of analysis. Contributions were separated when branches differed materially in their terminal objective, monitored quantity, stop/continue condition, update mechanism, output, assumptions, or guarantees. This decomposition preserves distinctions that would be lost if each study were represented as a single indivisible method.

The synthesis organizes the 95 contributions into five families defined by their terminal rationale: precision or uncertainty sufficiency, discrete-decision sufficiency, risk or loss sufficiency, operational-state sufficiency, and resource-aware acquisition and stopping. The documented evidence is then compared along five dimensions: finite-sample evidence, alternative-rule comparison, robustness, computational or scalability evidence, and implementation availability. These dimensions describe reported evidence and are not combined into a global readiness score.

The contributions of this survey are to:

- provide a transparent account of corpus formation;
- present a family-level anatomy of stopping mechanisms;
- summarize dimensional evidence profiles;
- provide contribution-level illustrations;
- offer conditional practitioner guidance;
- present a corpus-bounded analysis of research gaps and future directions.

Its purpose is not to identify a universally superior stopping rule, but to make explicit what different contributions monitor, why they terminate, what they return, what evidence supports them, and what conditions limit their use in another setting.

Defense constructive simulation provides a specific motivation for this comparison: operational analyses involve repeated scenario executions [7], making the decision about when sufficient evidence has been collected consequential for both analysis and computational effort. The survey therefore examines what the focal corpus documents about the evaluation of stopping criteria in this setting, while distinguishing statistical stopping across executions from the events that terminate an individual scenario.

Accordingly, the survey addresses five research questions:

- RQ1.** How do the retained contributions structure evidence acquisition and operationalize the data-dependent stop-or-continue decision?
- RQ2.** Which monitored quantities, assumptions, guarantees, and contribution-specific limitations define and qualify those stopping decisions?
- RQ3.** What evidence is reported for finite-sample behavior, comparison with alternative rules, robustness, computational or scalability behavior, and implementation availability?

TABLE 1. Positioning relative to earlier reviews and comparative studies. Publication years identify the earlier works, not their literature-search windows.

Work and type	Scope	Comparison unit	Evidence treatment
Law and Kelton (1982) [18]: survey and empirical comparison	Steady-state mean confidence intervals; sequentially determined run length	Sequential procedure	Coverage performance on stochastic models with known steady-state means
Law and Kelton (1984) [19]: survey and empirical comparison	Steady-state mean confidence intervals; run length fixed in advance	Fixed-sample procedure	Interval performance and sources of error on models with known steady-state means
Bicher et al. (2022) [9]: technical review and numerical evaluation	Replication counts for Monte Carlo mean estimation	Stopping-rule and algorithm combination	Tests of four combinations on an epidemic simulation model
This survey: mechanism and evidence synthesis	Data-dependent termination across different objectives; 2015-2025 focal corpus	Operational contribution, distinguishing variants	Reported evidence across five dimensions; no common-testbed reimplementation

RQ4. What does the focal corpus document about application settings, implementation, independent reuse, and transfer beyond the originating setting?

RQ5. Which cross-family and application gaps remain insufficiently supported within the focal corpus?

This article is organized as follows. Section II describes the survey methods, Section III presents the family and evidence synthesis, Section IV discusses validation gaps, their practical implications, survey limitations, and research directions, and Section V summarizes the conclusions.

II. METHODS

This section describes the survey design and scope, the search strategy used to identify the literature, and the procedures used to select studies and form the final corpus. Together, these steps define the evidential basis for the analyses presented in the subsequent sections.

A. SURVEY DESIGN AND SCOPE

This survey used a documented, protocol-driven systematic search and study-selection process to identify, select, and compare operational stopping criteria. A stopping criterion was considered operational when it used accumulating data to determine whether simulation, Monte Carlo, or sequential sampling effort should continue or terminate. PRISMA 2020 informed reporting of the following elements [20]:

- survey rationale and research questions;
- eligibility criteria;
- information sources and search strategies;
- study selection and corpus formation;
- contribution-level characteristics and qualitative synthesis; and
- interpretation, limitations, and supplementary-material availability.

PRISMA-S additionally informed database-search reporting [21].

The survey focused on studies published from 2015 through 2025, an inclusive eleven-calendar-year period. For conference papers, the temporal filter followed the year recorded in the indexed bibliographic record rather than the

conference-event year. This bounded period supported a detailed examination of contemporary stopping mechanisms, their empirical evaluation, and the conditions under which they were applied. The resulting corpus is therefore intended to characterize this defined period rather than the complete historical literature on stopping criteria.

A study was eligible for the primary corpus when it described an explicit data-dependent stop-or-continue decision, used information accumulated during the procedure, and addressed whether the simulation or sampling effort was sufficient. The stopping mechanism and the monitored quantity also had to be identifiable, and simulation or Monte Carlo had to be substantive to the decision rather than used only as an incidental evaluation tool.

Related areas were retained as background when they informed the theory, evaluation, or boundaries of the focal problem but did not satisfy the operational definition. This distinction allowed the survey to acknowledge adjacent sequential-stopping literatures without treating every use of termination, convergence, or adaptive sampling as an included stopping criterion.

Table 2 shows how the operational definition was applied to distinguish the primary scope from background, adjacent, and intentionally excluded literature.

The boundary assessment also clarified the treatment of the related areas identified during corpus assessment. MCMC stopping and Monte Carlo error control, simulation-output analysis, adaptive sampling and sequential experimental design, rare-event and reliability simulation, and operations-research applications were not excluded by domain label. Procedures from these areas entered the primary corpus whenever they satisfied the same operational eligibility criteria. Classical sequential-analysis studies and simulation-output literature outside the focal period were retained as theoretical or historical background when needed.

Statistical and stochastic curtailment provide a related but distinct form of data-dependent termination. Sie et al. applied stochastic curtailment to adaptive mastery testing, allowing an assessment to terminate when the probability of a different final classification became sufficiently small [22]. Although

TABLE 2. Operational treatment of literature at the focal boundaries

Boundary pattern	Treatment	Operational reason
Operational, data-dependent decision of sampling or simulation sufficiency	Primary scope	Uses accumulated information to decide whether further simulation or sampling is necessary.
Classical or pre-2015 methodological literature	Background foundation	Provides historical or theoretical foundations but falls outside the focal 2015-2025 period and primary corpus.
Convergence diagnostic without an operational sufficiency condition	Adjacent boundary	Indicates apparent stability but does not define when evidence acquisition should end.
Adaptive allocation or design without termination	Adjacent boundary	Determines where, what, or from which source to sample next, but not when total sampling is sufficient.
Curtailement of a trial or testing procedure	Adjacent boundary	Terminates a trial, test, or assessment rather than simulation or Monte Carlo sampling effort.
Surrogate-model refinement as the terminal objective	Intentionally outside	Determines when a surrogate or metamodel has been refined sufficiently, not whether simulation or sampling has produced sufficient inferential evidence.
Optimization-algorithm convergence	Intentionally outside	Terminates an optimization routine when its solution, objective value, or optimality gap stabilizes, rather than establishing simulation sufficiency.
Early stopping in trials, classification, or monitoring	Intentionally outside	Terminates participant recruitment, model training, classification, or process monitoring rather than simulation effort.
Simulation or Monte Carlo used only for evaluation	Intentionally outside	Uses simulation to test or compare another method under a fixed design; the rule does not determine how much simulation evidence to collect.

this mechanism shares the stop-or-continue logic examined here, it terminates an assessment rather than simulation or Monte Carlo effort. Curtailement was therefore retained as adjacent boundary literature rather than included in the primary corpus.

B. SEARCH STRATEGY

The search was conducted in Scopus and Web of Science on July 15, 2026. These databases were selected for their broad coverage of journals and conference proceedings across simulation, applied statistics, operations research, engineering, and computing. The search was restricted to English-language Articles and Conference or Proceedings Papers published from 2015 through 2025.

Four complementary search strategies were used to capture differences in how the literature describes data-dependent termination. The first targeted explicit stopping and termination terminology in simulation and sequential estimation. The second emphasized precision requirements, confidence-interval coverage, Monte Carlo error, and stopping-induced bias. The third focused on simulation output analysis and steady-state simulation, where stopping rules may be described through run length, batching, or fixed-width procedures. The fourth addressed the validation-and-performance question by targeting studies that evaluated, compared, benchmarked, or tested the finite-sample behavior, robustness, or stopping-induced bias of stopping procedures. This complementary route could retrieve evaluation-oriented records not captured through mechanism-focused terminology; records retrieved through it remained subject to the same eligibility criteria as records from the other strategies.

Table 3 presents the complete conceptual expression for

each strategy. The expressions were translated to the syntax of each database: searches were applied to titles, abstracts, and keywords in Scopus and to the Topic field in Web of Science; proximity was expressed using the corresponding database operator.

C. STUDY SELECTION AND CORPUS FORMATION

The database searches returned 364 occurrences. Records were matched using source identifiers, DOI information, and normalized titles and publication years. Deduplication removed 157 repeated occurrences, leaving 207 distinct records. 4 additional records were removed manually before scientific screening: 3 were residual duplicates, and 1 could not be located because all consulted links were unavailable. This left 203 records available for eligibility assessment.

Selection proceeded through title-and-abstract screening followed, when required, by full-text assessment. The eligibility criteria defined in Section II-A were applied to determine whether each record described an identifiable, data-dependent stopping decision substantively connected to simulation or Monte Carlo. This assessment excluded 130 records for content- or scope-based reasons. The main reasons were the absence of an operational stopping rule, a sample size or computational budget fixed in advance, adaptive allocation without a termination decision, a stopping objective outside the focal scope, incidental use of simulation or Monte Carlo, or insufficient information to identify the monitored quantity or stopping mechanism. Figure 2 presents the distribution of the exclusion reasons.

After scientific eligibility assessment, the authors applied a separate publisher boundary informed by an assessment of publisher signals related to archived Beall-list informa-

TABLE 3. Search strategies used to identify operational stopping criteria

Search focus	Complete conceptual search expression
Explicit stopping terminology	("stopping rule*" OR "stopping criterion" OR "stopping criteria" OR "sequential stopping" OR "termination rule*" OR "termination criterion") PROX/10 ("stochastic simulation" OR "Monte Carlo simulation" OR "simulation output analysis" OR "steady-state simulation" OR "steady state simulation" OR "sequential estimation" OR "sequential sampling")
Precision and sequential inference	("fixed-width" OR "fixed width" OR "absolute precision" OR "relative precision" OR "confidence interval width" OR "Monte Carlo standard error" OR "coverage probability" OR "stopping bias") AND ("sequential procedure*" OR "sequential estimation" OR "sequential sampling" OR "sequential stopping" OR multistage OR "multi-stage" OR "two-stage" OR "run length" OR batching) AND (simulation OR "Monte Carlo")
Simulation output analysis	("simulation output analysis" OR "steady-state simulation" OR "steady state simulation") AND ("stopping rule*" OR "stopping criterion" OR "sequential procedure*" OR "fixed-width" OR "fixed width" OR "run length" OR batching OR "batch means")
Evaluation and comparison	("stopping rule*" OR "stopping criterion" OR "stopping criteria" OR "sequential procedure*") AND (evaluation OR comparison OR benchmark* OR robustness OR "coverage probability" OR "finite sample" OR "stopping bias") AND ("stochastic simulation" OR "Monte Carlo simulation" OR "simulation output analysis" OR "steady-state simulation" OR "steady state simulation" OR "sequential estimation")

Note: Expressions were applied to titles, abstracts, and keywords through TITLE-ABS-KEY in Scopus and to the Topic field through TS in Web of Science. PROX/10 represents the corresponding proximity operator: W/10 in Scopus and NEAR/10 in Web of Science. All searches used the focal eligibility filters: publication years 2015-2025, English language, and Article or Conference/Proceedings Paper.

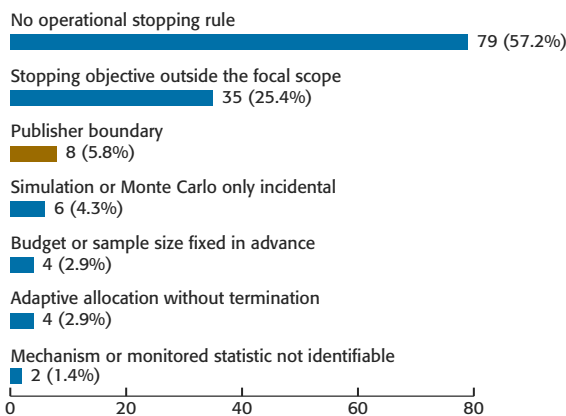


FIGURE 2. Reasons for exclusion from the focal corpus. An amber fill distinguishes the publisher boundary from content- or scope-based exclusions. Unavailable reports and operational removals are shown separately in Figure 3.

tion. This decision excluded 8 otherwise method-eligible records: 6 associated with MDPI, 1 with Frontiers, and 1 with Pushpa Publishing House. This boundary was applied separately from the scientific eligibility criteria and therefore does not constitute a judgment about the scientific validity of an individual study or the general quality of its publication venue. Figure 2 presents all 138 exclusions: 130 based on content or scope and 8 based on the separate publisher boundary. Full texts were sought for the remaining 65 reports.

11 reports remained identifiable through their bibliographic pages, but their full-text documents could not be obtained because access required subscriptions not covered by the authors' institutions and were therefore classified as unavailable rather than scientifically excluded. The final corpus comprised 54 primary studies. Figure 4 shows their distribution by publication year. The complete corpus-formation process is presented in Figure 3.

D. UNIT OF ANALYSIS

The document was the unit of study selection, whereas the operational contribution was the unit of analysis. A single study could report more than one stopping procedure, variant, or materially distinct decision rule.

To determine whether a procedure constituted a distinct contribution, its stopping objective, monitored quantity, continue-or-terminate condition, update mechanism, terminal output, assumptions, and stated guarantees were examined. Procedures were recorded separately when differences in these elements changed the operational decision or its interpretation. Parameter settings or experimental configurations of the same rule were not treated as separate contributions.

This process identified 95 operational contributions across the 54 primary studies. Each contribution retained its link to its source study and is represented separately in the contribution-level supplementary table *Operational Contributions and Evidence Profiles*, described in Section IV-F.

Evidence reported for one variant, branch, or application was not attributed to another contribution from the same study

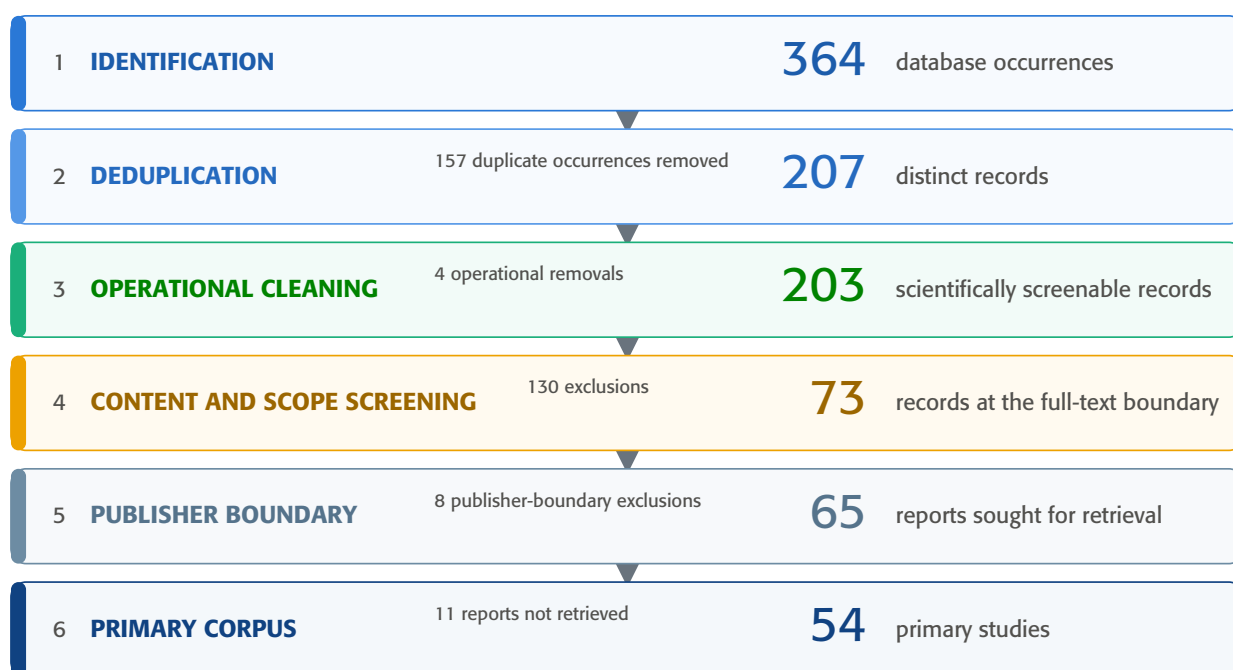


FIGURE 3. Focal corpus formation from 364 database occurrences to 54 primary studies. Operational removals, content- or scope-based exclusions, the publisher boundary, and unavailable reports represent distinct decision bases.

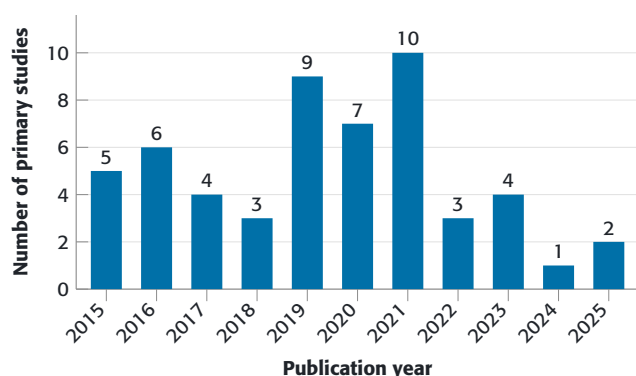


FIGURE 4. Annual distribution of the 54 primary studies in the focal corpus, 2015-2025. Each study is counted once according to its bibliographic publication year. Counts describe the selected corpus, not the overall publication volume of the field.

unless the source explicitly established that the evidence applied to both.

E. EVIDENCE APPRAISAL

Each contribution was appraised across five dimensions: finite-sample evidence, comparison with alternative stopping procedures, robustness analysis, computational or scalability evidence, and implementation availability. These dimensions describe the evidence documented for each contribution rather than the intrinsic quality of the method or whether its reported results were favorable.

Finite-sample evidence concerned direct examination of how the contribution behaved at stated finite sample sizes or computational budgets. Relevant outcomes included coverage, estimation error, decision accuracy, stopping effort, and resource cost under finite operating conditions. Asymptotic results alone did not constitute finite-sample evidence, nor did Monte Carlo simulation used only to derive or calibrate a procedure without evaluating its resulting operating behavior. Comparison evidence required another identifiable stopping or decision procedure to be evaluated under a common setting and outcome; changes in parameters, thresholds, or acquisition settings within the same rule were not treated automatically as comparisons with an alternative procedure.

Robustness evidence concerned directed examination of the contribution under perturbations of assumptions relevant to its validity or operation, including distributional shape, dependence, nonstationarity, and model misspecification. Routine variation of sample size, confidence level, tolerance, hyperparameters, or application scenarios was not considered a robustness analysis unless it deliberately tested such an assumption. Computational or scalability evidence concerned how resource measures such as runtime, memory, operations, evaluations, transmissions, or sample cost changed with problem size, budget, dimension, or workload. An isolated runtime, evaluation count, or qualitative complexity statement provided computational information but did not by itself demonstrate scaling behavior.

Implementation availability concerned the material pro-

vided to execute the specific contribution, rather than the completeness of its mathematical description. Demonstrated availability required a locatable and usable artifact, such as source code, a software package, an executable supplement, or a reproducible procedure specified sufficiently for use. Limited availability encompassed heterogeneous forms of partial support, including pseudocode or algorithmic steps without executable code, a named software environment without the complete evaluated implementation, code available only on request, incomplete or version-ambiguous repository material, and artifacts that omitted branches or execution-critical conventions. The limited state therefore represents partial reproducibility, not a uniform level of software availability.

The evidence profile uses three states: demonstrated evidence (✓), limited or partial evidence (∼), and evidence not reported (—). A demonstrated state indicates that the dimension was substantively examined, even when the reported outcome was unfavorable. A limited state indicates relevant but restricted evidence, such as evaluation of only some branches, assumptions, settings, or scales. A not-reported state indicates that the study did not present evidence addressing that dimension for the specific contribution. It does not imply that the procedure would perform poorly if evaluated.

The five dimensions were retained as a multidimensional evidence profile and were not combined into a readiness score, ranking, or universal recommendation.

F. SYNTHESIS

The contributions were synthesized through structured comparison of their terminal rationale and operational mechanism. The comparison considered what condition was interpreted as sufficient for termination, the quantity monitored, how information was acquired and updated, the continue-or-terminate condition, and the output returned when the procedure stopped. This synthesis produced five families: precision or uncertainty sufficiency, discrete-decision sufficiency, risk or loss sufficiency, operational-state sufficiency, and resource-aware acquisition and stopping.

Each contribution was assigned to one family according to its dominant terminal rationale. When a contribution contained multiple branches or successive stopping decisions, their distinctions and evidential limits were retained at the contribution level rather than representing the contribution in more than one family. Family membership therefore indicates a shared operational rationale, not equivalence of assumptions, guarantees, or empirical support.

Evidence states were subsequently aggregated by family for each of the five appraisal dimensions. The denominator for each family-level profile was the number of operational contributions assigned to that family. Because a study could contain multiple contributions, these aggregates characterize the mapped procedures and should not be interpreted as counts of independent studies. Contribution and distinct-study totals are therefore reported separately in the results.

G. FORWARD-CITATION ASSESSMENT

To examine evidence beyond the focal corpus without expanding it, a directed forward-citation assessment covered 53 primary studies in Scopus. The remaining study [23] was searched in Web of Science and had no citing documents. The assessment therefore covered all 54 primary studies.

The Scopus searches returned 458 unique citing documents. We removed 20 documents already included in the 54-study primary corpus, leaving 438 external candidates. Title-and-abstract screening assessed whether each candidate indicated use, evaluation, comparison, adaptation, or application of a stopping procedure described in the cited primary study. Figure 5 summarizes the screening and the subsequent cross-classification of the 68 potentially relevant external documents by indication status and detected authorship overlap.

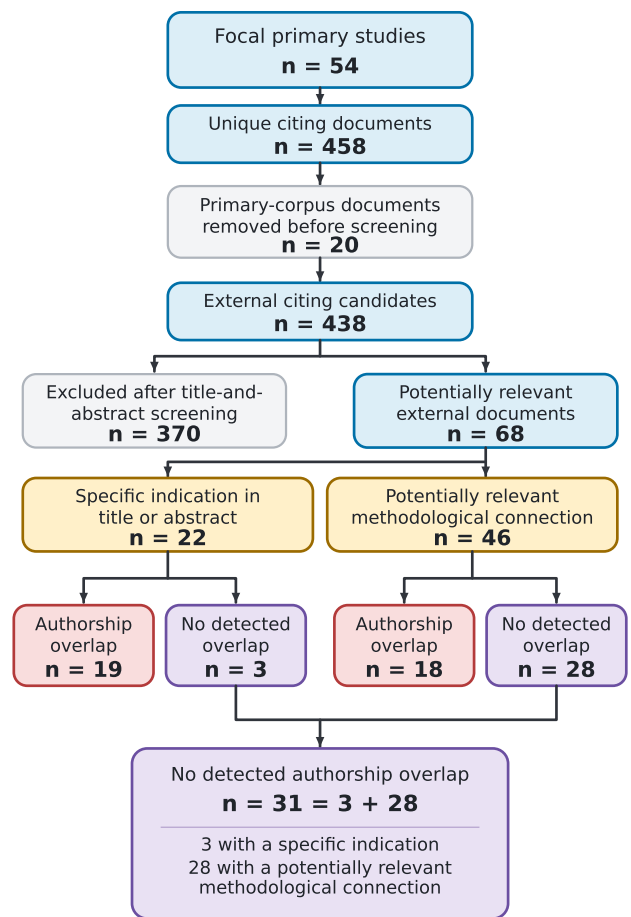


FIGURE 5. Forward-citation assessment of the 54 focal primary studies. After documents already in the primary corpus were removed, the 68 potentially relevant external citing documents were cross-classified by indication status and detected authorship overlap; the two no-overlap branches converge on the 31 documents without detected overlap.

Authorship was compared between each retained external citing document and every primary study through which it entered the assessment. A citing document was conservatively classified as having overlapping authorship when it shared at least one author with any corresponding primary study.

Comparison of complete Scopus Author ID lists identified 34 such documents. The authors adjudicated 3 additional exact-name cases in which affiliation changes were associated with separate Scopus profiles and confirmed that they represented the same authors. Together, the 34 Author ID matches and 3 adjudicated exact-name cases yielded 37 documents with overlapping authorship, which were treated as evidence of continuity involving an originating group rather than as candidates for independent evaluation. The remaining 31 had no detected authorship overlap: 3 contained a target-specific indication of use or evaluation, while 28 indicated a potentially relevant methodological connection whose exact procedural variant and role remained unresolved from the title and abstract. Candidate full texts were not examined.

III. TAXONOMY AND EVIDENCE PROFILES OF SIMULATION STOPPING CRITERIA

The taxonomy separates stopping criteria by what termination is intended to establish, rather than by acquisition architecture alone. The evidence profiles reveal uneven support across dimensions, with finite-sample behavior and alternative comparisons documented more often than robustness and complete implementation availability.

A. FAMILY ANATOMY

This subsection addresses RQ1 and RQ2. Table 4 compares the five families through a common operational anatomy: the quantity or state monitored, the acquisition or update process, the condition interpreted as sufficient for termination, the terminal output, and the principal boundary on interpretation. The families describe shared terminal rationales rather than interchangeable procedures. Assumptions, guarantees, and empirical evidence remain attached to the individual contribution and, where applicable, to its specific branch.

Precision or uncertainty sufficiency comprises 65 contributions, organized into four subfamilies according to the inferential target: mean estimation (23), estimation of other scalar parameters (25), quantile estimation (8), and joint parameter estimation (7). Only Family 1 is disaggregated because its concentration of 65 of the 95 contributions required a more informative level of granularity; no equivalent subdivision with defined boundaries and reconciled counts was established for the smaller families. Mean estimation includes individual means, differences of means, and common or projected means; other scalar parameters include variances, correlations, risk and odds ratios, and population sizes. Joint parameter estimation concerns a set of parameters considered together in the stopping decision, including vectors of means. Two further contributions from Flegal and Gong [41] provide procedures for means, quantiles, and joint estimation and are therefore reported separately as cross-cutting contributions.

Across the precision subfamilies, sequential monitoring updates uncertainty until the target is reached, whereas staged procedures use pilot or interim information to determine the remaining sampling effort. First-passage monitoring, batch-

ing, and staged sample-size determination retain distinct validity conditions despite their shared precision rationale.

For discrete-decision sufficiency, crossing an evidential boundary must be distinguished from a cap-forced decision: operational termination does not establish that the regular sufficiency condition was attained.

Risk or loss sufficiency accounts for the consequences or costs of estimation and decision error, rather than precision alone. Guarantees remain specific to the loss function, prior or parametric model, limiting regime, and procedural branch; reduced sampling effort does not establish risk optimality.

Operational-state sufficiency groups heterogeneous diagnostic, structural, and control gates that authorize termination or transition. Their shared operational role does not imply a common mathematical guarantee.

Resource-aware acquisition and stopping incorporates resource consumption into acquisition, termination, or both. Reaching an information target differs from reaching a budget, time, or computational boundary: the latter limits consumption without establishing statistical sufficiency.

B. EVIDENCE PROFILES

This subsection addresses RQ3. Table 5 summarizes the five evidence dimensions defined in Section II-E by family; Figure 6 presents the overall distribution across the 95 contributions. Counts retain the unequal family denominators (4-65 contributions), avoiding misleading percentage comparisons. Contribution-level branches, guarantees, evidence states, and limitations are documented in the supplementary table *Operational Contributions and Evidence Profiles*.

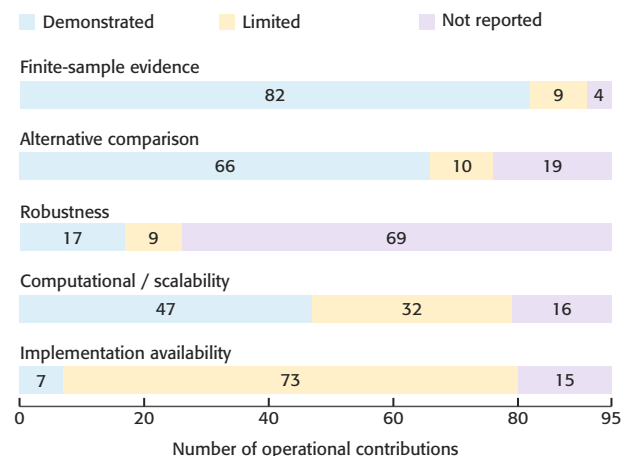


FIGURE 6. Evidence profiles across the 95 operational contributions. Finite-sample evaluation and comparison with alternative procedures predominate, whereas robustness is less frequently examined and implementation support is mostly partial.

C. ILLUSTRATIVE CONTRIBUTIONS

The examples illustrate within-family variation without implying representativeness, superiority, or recommendation. The tables provide at-a-glance comparisons of acquisition and

TABLE 4. Operational anatomy of the five stopping-criterion families

Family	Counts	Contributing studies	Monitored quantity or state	Acquisition, update, and terminal condition	Typical output
Family 1. Precision or uncertainty sufficiency	65 / 38	[5], [23]–[59]	Interval width, standard error, confidence-region extent, or another measure of uncertainty or precision.	Update the estimate and its uncertainty sequentially or in stages. Terminate when the prescribed precision is reached, or use pilot or interim estimates to determine the required sample size or run length.	Estimate accompanied by an interval, region, or precision statement.
Family 2. Discrete-decision sufficiency	5 / 3	[60]–[62]	Posterior loss or probability, detection statistic, decision zone, or evidence separating alternatives.	Update the evidence until a boundary separates classes, hypotheses, or alternatives. Some procedures impose a cap that forces a terminal decision.	Classification, detection result, or selected alternative or set.
Family 3. Risk or loss sufficiency	12 / 11	[44], [48], [50]–[53], [56], [63]–[66]	Estimated risk, posterior expected loss, regret, or cost-weighted estimation error.	Update the risk estimate or the required sampling effort until the relevant loss or objective that balances cost and error is satisfied.	Risk-oriented point estimate or terminal sampling decision.
Family 4. Operational-state sufficiency	9 / 6	[29], [67]–[71]	Agreement, error bound, validation score, relative error, structural persistence, or controller state.	Continue diagnostic, structural, or control updates until the operational state passes its specified gate.	Stable or validated estimate, identified structure, ranking, sensitivity measure, or completed control action.
Family 5. Resource-aware acquisition and stopping	4 / 4	[72]–[75]	Information or precision proxy together with evaluations, transmissions, time, budget, or predicted computational burden.	Prioritize informative acquisitions and terminate when the information target is reached or an explicit resource boundary prevents further acquisition.	Estimate, selected solution, reliability quantity, or completed acquisition process.

Note: Counts refer to operational contributions / distinct primary studies. Study counts are not additive because a study may contribute to more than one family.

stopping, whereas the accompanying prose retains the technical mechanisms, outputs, assumptions, guarantees, results, and limitations that the summaries cannot capture.

1) Illustrative Precision and Uncertainty Contributions

Table 6 contrasts three precision or uncertainty contributions with different acquisition and stopping patterns.

TABLE 6. Acquisition patterns illustrated within the precision or uncertainty family.

Illustrative contribution	Key distinctions in sampling and stopping
Zoubeidi and El-Bassiouni [59]	Each new cluster updates the information used to assess the confidence-region stopping condition.
Mukhopadhyay and Hu [48]	The pilot determines the required total sample size; sampling then ends immediately or after one additional batch.
Alexopoulos et al. [24]	Blockwise diagnostics prepare the data before the corrected interval is checked against the precision target.

Note: These examples contrast acquisition patterns, not the strength of their guarantees.

Zoubeidi and El-Bassiouni [59] considered sequential construction of a fixed-size confidence region for the parameters of a mixed-effects logistic regression model. After each newly acquired cluster, the model is refitted and the estimated information matrix is updated. The confidence ellipsoid is scaled

to have the prescribed longest-axis width, and sampling terminates when the smallest eigenvalue of the estimated information matrix satisfies the corresponding moderated chi-square information threshold. The procedure returns a confidence region for the parameter vector or a specified linear transformation of it. The coverage guarantee is asymptotic as the requested width tends to zero and depends on regularity conditions and a valid positive-definite information estimate; the reported finite-sample coverage was not uniformly nominal.

A different acquisition pattern appears in the two-stage procedure of Mukhopadhyay and Hu [48]. An initial sample estimates the unknown dispersion using the sample variance or finite-sample-corrected squared forms of Gini's mean difference and mean absolute deviation. That estimate determines the total sample size required for a confidence interval with the prescribed half-width. The procedure does not repeatedly inspect a precision boundary after the pilot. Its reported guarantees are asymptotic under independent normal sampling and a valid known lower bound on the standard deviation. Simulations covered several target sample sizes, but coverage was below the nominal level in some of the smallest settings.

Sequest, proposed by Alexopoulos et al. [24], illustrates a blockwise simulation-output procedure for estimating a stationary quantile from a dependent output sequence. The procedure increases the batch size and run length, removes warm-up observations, and updates correlation- and skewness-adjusted confidence limits. Its terminal precision branch stops when the corrected half-width is below an absolute target or

TABLE 5. Evidence profiles of the five stopping-criterion families

Family	Contributions	Finite-sample evidence	Alternative comparison	Robustness	Computational / scalability	Implementation availability
Precision or uncertainty sufficiency	65	✓ 58; ~ 5; — 2	✓ 45; ~ 8; — 12	✓ 14; ~ 6; — 45	✓ 33; ~ 20; — 12	✓ 5; ~ 50; — 10
Mean estimation	23	✓ 21; ~ 1; — 1	✓ 17; ~ 2; — 4	✓ 7; ~ 0; — 16	✓ 11; ~ 11; — 1	✓ 2; ~ 16; — 5
Other scalar-parameter estimation	25	✓ 24; ~ 0; — 1	✓ 17; ~ 3; — 5	✓ 1; ~ 4; — 20	✓ 13; ~ 5; — 7	✓ 0; ~ 21; — 4
Quantile estimation	8	✓ 4; ~ 4; — 0	✓ 5; ~ 2; — 1	✓ 6; ~ 0; — 2	✓ 6; ~ 2; — 0	✓ 3; ~ 5; — 0
Joint parameter estimation	7	✓ 7; ~ 0; — 0	✓ 4; ~ 1; — 2	✓ 0; ~ 0; — 7	✓ 3; ~ 0; — 4	✓ 0; ~ 6; — 1
Cross-cutting contributions	2	✓ 2; ~ 0; — 0	✓ 2; ~ 0; — 0	✓ 0; ~ 2; — 0	✓ 0; ~ 2; — 0	✓ 0; ~ 2; — 0
Discrete-decision sufficiency	5	✓ 4; ~ 0; — 1	✓ 3; ~ 0; — 2	✓ 1; ~ 0; — 4	✓ 3; ~ 2; — 0	✓ 2; ~ 3; — 0
Risk or loss sufficiency	12	✓ 11; ~ 1; — 0	✓ 9; ~ 1; — 2	✓ 2; ~ 0; — 10	✓ 6; ~ 4; — 2	✓ 0; ~ 8; — 4
Operational-state sufficiency	9	✓ 6; ~ 2; — 1	✓ 5; ~ 1; — 3	✓ 0; ~ 2; — 7	✓ 2; ~ 5; — 2	✓ 0; ~ 9; — 0
Resource-aware acquisition and stopping	4	✓ 3; ~ 1; — 0	✓ 4; ~ 0; — 0	✓ 0; ~ 1; — 3	✓ 3; ~ 1; — 0	✓ 0; ~ 3; — 1
All contributions	95	✓ 82; ~ 9; — 4	✓ 66; ~ 10; — 19	✓ 17; ~ 9; — 69	✓ 47; ~ 32; — 16	✓ 7; ~ 73; — 15

Note: For each row and dimension, counts sum to the number of contributions: ✓ demonstrated evidence; ~ limited or partial evidence; — evidence not reported. The five indented rows partition Family 1 and are not counted again in the overall total. Aggregation does not imply that evidence applies to every contribution in a family.

TABLE 7. Sampling and stopping patterns illustrated within the discrete-decision family.

Illustrative contribution	Key distinctions in sampling and stopping
Batur and Choobineh [60] - QSE	Sampling stops when one alternative remains or all remaining pairs are declared equivalent; the output may therefore be a set.
Cui et al. [61]	Each radar scan updates posterior losses. Acquisition stops at a decision threshold, or a maximum scan count forces a decision.
Solanky and Zhou [62] - sequential procedure	Each round samples every population. Sampling stops when the accumulated sample size reaches the requirement calculated from the updated variance estimate.

Note: A decision forced by a sampling cap does not necessarily satisfy the regular evidential stopping condition. Suffixes identify the specific contribution when a study reports more than one operationally distinct procedure.

below a prescribed fraction of the estimated quantile. The procedure also contains internal checks of batch behavior, dependence, and skewness. The underlying interval theory is asymptotic under dependence and local-density conditions and does not establish a finite guarantee for the complete adaptive path. In the reported experiments, the relative-precision branch was examined prospectively, whereas the absolute-precision branch was not tabulated.

2) Illustrative Discrete-Decision Contributions

Unlike precision-oriented termination, discrete-decision sufficiency ends in a categorical output. Table 7 compares three of the family's five contributions, representing all three source studies. Their outputs are a set of acceptable alternatives, a declaration of target presence, or population classes.

Batur and Choobineh [60] developed sequential procedures for selecting alternatives according to an estimated quantile. Their Quantile-Based Selection with Equivalency procedure updates pairwise quantile differences and paired variance estimates while eliminating dominated alternatives and suspending acquisition for alternatives whose necessary pairwise relations have already been resolved. Elimination and completion of individual pairwise relations are internal decision gates, whereas termination occurs only when the global condition on the remaining set is satisfied. The reported selection guarantees are asymptotic as the indifference-zone width tends to zero; the finite experiments support observed selection performance but do not establish an exact finite-sample probability of correct selection.

Cui et al. [61] considered sequential track-before-detect processing in which each additional radar scan updates posterior losses for declaring target absence or presence. These losses can also incorporate errors in the estimated final position and trajectory. A declaration of presence is accompanied by position and trajectory estimates. Despite these accompanying estimates, presence or absence remains the decision that ends acquisition. Such a cap guarantees operational termination but does not demonstrate that the regular evidential condition was attained, and the probability bounds derived for the regular boundaries do not automatically certify the forced-decision branch.

Solanky and Zhou [62] addressed classification of treatments relative to a control into bad, intermediate, and good groups under a normal model with unknown common variance. The sequential procedure begins with observations from every treatment and the control, updates a pooled variance estimate, and continues with one new observation from every population at each round. Sampling stops globally at the first sample size that is at least the size required by the current variance estimate; the observed mean differences are then

compared with the classification boundaries to assign the treatments to the three groups. Accumulated variance information determines when the data are considered sufficient, whereas the mean differences determine the terminal class labels. The procedure terminates almost surely under the stated model, but its principal efficiency and probability-of-correct-decision results are asymptotic as the indifference zones narrow. They depend on independent normal populations, a common variance, symmetric indifference zones, and the configuration used in the formal analysis.

3) Illustrative Risk and Loss Contributions

Table 8 compares three risk or loss contributions whose terminal objectives use a frequentist risk approximation, a loss combining estimation error and sampling cost, or Bayesian posterior expected loss.

TABLE 8. Sampling and stopping patterns illustrated within the risk or loss family.

Illustrative contribution	Key distinctions in sampling and stopping
Karmakar and Mukhopadhyay [63]	Each observation updates an estimate of the risk-optimal sample size. Sampling stops when the accumulated size reaches that estimate.
Mukhopadhyay and Hu [48]	The pilot determines a total sample size balancing estimation error and sampling cost; sampling ends after the pilot or one additional batch.
Mukherjee et al. [65]	Fisher information guides source selection, while posterior expected loss determines when acquisition stops.

Note: The procedures use different loss formulations and assumptions; their guarantees are not interchangeable.

Karmakar and Mukhopadhyay [63] developed a sequential procedure for estimating the fixed coefficient component of a multivariate random-coefficient autoregressive process. After an initial sample, each new observation updates the coefficient estimates, residual moments, and nuisance quantities entering an approximation to predictive risk. The boundary targets a balance between conditional-predictor estimation error and sampling cost. The procedure is shown to terminate almost surely under its stated conditions and becomes asymptotically equivalent to the derived oracle as the cost parameter enters the relevant limiting regime. That oracle minimizes an asymptotic risk approximation, however, rather than an exact finite-sample risk. The reported simulations compare the sequential procedure with the oracle across several orders, dimensions, and reported dispersion configurations, but do not establish robustness to model misspecification or provide a complete treatment of singular plug-in states.

Mukhopadhyay and Hu [48] also used a two-stage architecture for minimum-risk point estimation of a normal mean. The pilot estimates the unknown dispersion using finite-sample-corrected forms of the sample standard deviation, Gini's mean

difference, or mean absolute deviation. This estimate determines a total sample size that balances squared estimation error against a cost assigned to each observation. Sampling stops after the pilot if that total has already been reached; otherwise, one additional batch completes the calculated requirement, and the terminal output is the sample mean. This is the same broad acquisition architecture used by the study's fixed-width confidence-interval contribution, but the terminal objective balances estimation error and sampling cost rather than satisfying a prescribed precision target. The minimum-risk guarantees are asymptotic as the sampling cost tends to zero under independent normal sampling and a valid known lower bound on the standard deviation. Simulations showed risk efficiency approaching its target in the examined settings, but did not test nonnormality, contamination, misspecification of the lower bound, or costs that vary by observation or stage.

Mukherjee et al. [65] considered estimation of a shared parameter from multiple data sources. At each step, the source with the greatest Fisher information at the current posterior mean is selected, a new observation is acquired, and the posterior distribution is updated. The intended terminal condition is the first time that posterior expected squared loss falls below the prescribed threshold, after which the posterior mean is returned. The stated optimality result is asymptotic in the authors' limiting regime and does not provide a finite-threshold guarantee of accuracy or termination. Moreover, the reported numerical example always favors the same source, so it compares the resulting procedure with alternatives but does not empirically exercise the adaptive source-switching behavior that motivates the multimodal formulation.

4) Illustrative Operational-State Contributions

Table 9 compares four operational-state contributions whose gates use estimated error, diagnostic agreement, structural persistence, validation status, or the state of a control action.

Wu et al. [71] used Monte Carlo simulation to estimate mean trajectory error for an underwater glider operating under uncertain ocean currents. Each replication produces a complete simulated trajectory, after which the running mean and sample variance are updated. The published rule does not provide a finite sequential coverage guarantee. The reported simulations stopped at 1,950 replications without feedback and 2,070 with feedback. However, the first eligible inspection, a maximum replication count, behavior near a zero running mean, and adjustment for repeated inspection were not reported.

Schmalofski et al. [69] considered sequential atomistic simulations for mapping grain-boundary energy landscapes and identifying minima, valleys, and cusps. After each acquisition, a Kriging model of the energy landscape is refitted, and candidate locations are selected according to estimated interpolation uncertainty. This persistence does not prove that every relevant structure has been found, because the result remains conditional on the sampling policy, discretization, surrogate, and cusp detector. The reported one- and two-

TABLE 9. Operational checks and stopping patterns illustrated within the operational-state family.

Illustrative contribution	Key distinctions in sampling and stopping
Wu et al. [71] - Monte Carlo estimation	Every 30 replications, the relative confidence half-width is checked against the prescribed tolerance.
Schmalofski et al. [69]	Acquisition stops after three consecutive iterations with sufficiently small surface changes and an unchanged number of detected cusps.
Tavana et al. [70]	Entropy-based validation governs whether additional evaluations are extrapolated. Continuation requires a decrease in the global entropy measure and in the entropy products of alternatives whose ranking positions change.
Wu et al. [71] - feedback duration	The controller uses yaw error to determine a corrective action and its duration before restoring nominal control.

Note: The mechanisms govern different processes: acquiring simulation output, extrapolating evaluations, or ending a control action. They do not share a common convergence guarantee.

dimensional material cases included a complex 2D case in which sequential acquisition identified structures missed by an equal-budget regular grid, but no probabilistic guarantee of complete discovery, independent repetitions, or demonstrated result for the full five-degree-of-freedom space was provided.

Tavana et al. [70] embedded an entropy-based gate in a decision-support framework for resolving uncertainty in expert evaluations. The framework extrapolates additional evaluations from those initially provided by experts. Candidate rerankings are assessed through global and local entropy changes. A reranking passes the validation gate only when the framework's global product of marginal entropy sums decreases and each repositioned alternative also satisfies its local entropy condition. This data-dependent gate controls continuation of the extrapolation process: validation permits another loop, whereas failed validation leads to the termination branch shown in the framework's flowchart, which also includes a domain-modification option. Passing the gate establishes internal consistency under the framework's entropy criteria, not correctness of the resulting ranking, convergence of repeated loops, or agreement with an external decision truth. In the reported NATO-enlargement analysis, the global entropy measure decreased within the initial-evaluation-constrained domain, but the local entropy product increased for Bosnia and Herzegovina. The authors therefore required its relative ranking position to remain unchanged while validating the resulting ranking. The study did not provide a convergence proof, a maximum number of loops, or an externally validated target ranking.

A second contribution from Wu et al. [71] concerns the duration of a temporary feedback action rather than the number of Monte Carlo replications. This is not a statistical conver-

gence rule. Simulation results reported lower mean trajectory error with feedback than without it for the selected configuration. Nevertheless, the rule is model-derived and heuristic; no formal guarantee of global stability or termination was established, and saturation, ramp behavior, zero-denominator states, and the flowchart's global end condition were not fully specified.

5) Illustrative Resource-Aware Contributions

Table 10 compares all four resource-aware contributions, whose acquisition policies, terminal conditions, or both account for evaluations, transmissions, time, or computational effort.

TABLE 10. Sampling and stopping patterns illustrated within the resource-aware family.

Illustrative contribution	Key distinctions in sampling and stopping
Semelhago et al. [72]	Alternatives are evaluated in parallel batches. Search may stop at an estimated optimality-gap threshold or a time or iteration limit.
Dang et al. [73]	Posterior uncertainty guides the next expensive evaluation; evaluations stop when a posterior coefficient-of-variation bound for one augmented scalar component meets its tolerance.
Quan et al. [74]	Already measured sensor data are transmitted in order of known precision. Further transmissions are intended to stop when the posterior interval-width target is met.
Zhao et al. [75]	Adaptation stops when predicted total computational burden no longer decreases. A separate coefficient-of-variation condition ends the subsequent estimation phase.

Note: The procedures control different resources. Reaching a resource limit does not establish that the statistical target has been attained.

Semelhago et al. [72] considered simulation optimization over a finite set of alternatives through a parallel version of the Rapid Gaussian Markov Improvement Algorithm. A Gaussian Markov random field is updated from completed simulation batches, and a greedy q -complete-expected-improvement heuristic selects q candidates, which are evaluated synchronously with the incumbent. Crossing the estimated precision threshold represents the procedure's information-based stopping condition, while exhausting a budget only establishes that no further resources will be spent. The study compared alternative batch-selection strategies over 50 macro-replications and measured the cost of constructing the batches, but did not directly evaluate the reliability of the terminal precision decision. Batch construction is heuristic, and its computational cost increased sharply with batch and candidate-set size.

Dang et al. [73] developed an active-learning procedure for estimating a failure-probability function when parameters of

the input distributions are imprecisely known. A Gaussian-process surrogate is fitted in a joint random and epistemic space. At each iteration, the next point is selected by maximizing its upper-bound posterior variance contribution, the expensive limit-state function is evaluated there, and the surrogate is updated. The terminal gate controls a model-based posterior bound for one augmented scalar component, not pointwise or uniform error of the full failure-probability function, surrogate misspecification, finite candidate pools, or omitted higher-order components. Three structural-reliability examples reported substantially fewer expensive evaluations than selected alternatives, although in the third example the external comparison covered only the constant component. The procedure specified neither a mandatory consecutive-pass safeguard nor a cap or pool-exhaustion policy if the tolerance was not attained.

Quan et al. [74] considered sequential estimation in a sensor network where measurements have already been taken but need not all be transmitted to a fusion centre. After each received measurement, the fusion centre updates the conjugate Gaussian posterior and accumulates the corresponding measurement precision. The remaining transmissions are intended to be cancelled when that accumulated precision is sufficient to attain the prescribed posterior interval width, although the cancellation mechanism is not specified. The stopped resource is communication. Under the stated Gaussian model, crossing the threshold establishes the model-based posterior-width condition. It does not establish frequentist coverage or account for failure to reach the threshold with the available sensors. Synthetic experiments compared ordered, random, and all-sensor transmission using mean squared error and transmission counts. The reported resource measure was the number of transmissions, however, not total energy; sensing, listening, synchronization, collision, cancellation, and metadata costs were not measured.

Zhao et al. [75] combined cross-entropy adaptation, subset simulation, Metropolis-Hastings sampling, and importance sampling for composite power-system reliability evaluation. During adaptation, when its burden predictor is calculable, the procedure estimates total computational burden as the sum of the cost already incurred and the predicted cost of the subsequent importance-sampling phase. Under the pseudocode convention, adaptation stops at the first iteration that no longer reduces this total; the newest parameter set is discarded and the preceding importance-sampling distribution is retained. Resource economy and statistical sufficiency are therefore addressed at different stages of the operational pipeline. The first observed non-decrease is not a demonstrated global minimum of computational burden, and the prediction uses assumptions that are not fully aligned with every implemented sampling component. Comparisons with alternative reliability methods reported lower runtime and normalized burden in several power-system cases, but no independent repetitions or uncertainty intervals were provided, and the measured cost remains dependent on platform and implementation.

IV. VALIDATION GAPS AND RESEARCH DIRECTIONS

The discussion examines the principal validation gaps, their implications for practice, the limitations of the survey, and priorities for further research.

A. RESEARCH GAPS

This subsection addresses RQ5. Two cross-family gaps concern the evidential basis and operating conditions of current stopping criteria. A separate, corpus-bounded application gap in defense constructive simulation is examined in Section IV-B.

1) External Validation and Transferability

This subsection, together with the application analysis in Section IV-B, addresses RQ4. Evidence produced within the originating study establishes the initial theoretical or empirical support for a stopping procedure, but provides limited information about its reproducibility under independent implementation or its validity beyond the conditions in which it was developed. The relations identified within the 54-study focal corpus consisted predominantly of continuations by the originating research group or instances of reuse for which the operational variant could not be determined precisely. None contained sufficient documentation to confirm independent evaluation of the same variant or its transfer to a materially different domain. This result indicates unresolved evidence rather than the absence of subsequent validation.

The external assessment likewise identified candidate signals rather than confirmed external validation. Because the full texts and relationships between research groups were not examined, these signals do not establish independent validation. They also do not determine whether the same procedural variant was evaluated, which properties were tested, or whether the application domain differed materially from the original setting.

Subsequent studies do not consistently document their relationship to the originating contribution, making it difficult to distinguish conceptual reuse from independent implementation or a new application context from substantive evidence of transferability.

2) Stopping under Nonstationary and Heterogeneous Conditions

Many stopping criteria rely on assumptions of independent observations, stationarity, or a known dependence structure. These assumptions permit tractable analysis, but restrict the conditions under which the resulting guarantees can be interpreted. In operational settings, the statistical properties of the observed stream may vary during acquisition: dependence may change, observations may originate from heterogeneous operating conditions, and transitions between regimes may alter the information provided by subsequent samples. A criterion calibrated under a stable process may then interpret a transient period of apparent stability as sufficient evidence.

17 of the 95 contributions provided demonstrated robustness evidence. Illustrative studies examined specific depar-

tures from idealized sampling, including contrasting dependence structures and nonnormal marginal distributions [28], and nonnormal populations generated from gamma and log-normal distributions [35]. However, these analyses generally considered predefined forms of dependence or isolated perturbations. The corpus provides limited evidence about stopping when dependence, heterogeneity, and changes in the data-generating process occur jointly or are not known in advance. The survey therefore provides limited information about when existing guarantees remain valid as these conditions change.

B. DEFENSE AND OPERATIONAL SIMULATION AS AN UNRESOLVED APPLICATION GAP

Among the 95 retained operational contributions, none was evaluated in a defense constructive-simulation or military operational-analysis workflow. This result concerns the evaluations documented in the 54-study focal corpus; it does not imply that methods developed in civilian, engineering, or dual-use settings could not be adapted to defense applications. It identifies a mismatch between the operational demands documented in defense simulation and the settings in which statistically motivated stopping contributions have been evaluated.

The defense-simulation platform landscape includes multiple commercial and government environments. The comparison accompanying AsaPy examines seven commercial computer-generated-forces packages (STAGE, VR-Forces, SWORD, VBS4, DirectCGF, Steel Beasts Pro, and FLAMES) and separately discusses the government-off-the-shelf AFSIM environment [76]. An independent review provides a separate comparison of FLAMES, XSIM, AFSIM, and E-CARGO in terms of distributed capability, modularity, scalability, and generality [77]. Because the authors could not obtain FLAMES, AFSIM, or E-CARGO, those assessments relied on public information released by the software developers rather than direct experimentation. These systems support different combinations of scenario construction and execution, synthetic-force representation, training, operational analysis, and experiment preparation. Among them, FLAMES and AFSIM are identified as the closest to AsaPy in their treatment of experimental design and simulation data, although the reviewed capabilities remain oriented toward experiment configuration, recording, visualization, or externally supported analysis rather than an integrated statistical stopping procedure [76]. The platforms are therefore cited as operational context, not as evidence that they implement statistically justified replication stopping or as additions to the 95-contribution focal corpus.

The Aerospace Simulation Environment (ASA) demonstrates the scale and structure of this operational demand. ASA is a distributed constructive-simulation environment for military operational scenarios in which analysts define scenario templates, vary initial conditions, execute repeated scenarios across networked nodes, retain selected messages and performance metrics, and analyze the results for opera-

tional planning and decision support [7]. The environment has supported studies involving thousands of engagements and tens of thousands of simulated weapon launches [7]. These workflows make the number and cost of scenario executions operationally consequential, but the published ASA description does not itself define a statistical rule for determining when the accumulated executions are sufficient.

The companion library AsaPy addresses batch continuation more directly. Its execution controller divides a planned batch into successive chunks, executes one chunk at a time, evaluates selected output metrics, and invokes a user-defined function to decide whether to stop or launch the next chunk [76]. The published illustration compares the expected value of a selected metric before and after the latest chunk and stops when the variation falls below a threshold; the illustration is explicitly described as naive [76]. The article does not specify a confidence or coverage guarantee, calibrate the threshold against premature stopping, provide a joint rule for multiple outputs, or compare the stopping decision with completion of the planned batch. AsaPy therefore documents an operationally integrated, data-dependent batch-control mechanism, but not a statistically validated stopping criterion.

A separate distinction concerns the endpoint of each simulated scenario. Adjacent defense-modeling research examines when a squad-level combat engagement should end through conditions such as total force incapacitation, retreat triggers, leadership loss, and resource exhaustion [78]. These state- and event-based rules define the terminal semantics of an individual combat scenario, and their selection can alter modeled victory and casualty outcomes [78]. They do not determine when repeated stochastic executions have accumulated sufficient evidence for estimation or decision support. Scenario termination and replication-process stopping are therefore related operational concerns, but they address different stopped processes.

The closest contributions inside the focal corpus do not close this gap. One contribution evaluated a sequential track-before-detect procedure in a synthetic single-target radar model, where the stopped acquisition unit is the radar scan rather than a batch of constructive operational scenarios [61]. Another evaluated an entropy-based continuation gate through a NATO-enlargement ranking problem, but the stopped process generates additional subjective expert evaluations rather than military simulation replications [70]. These studies demonstrate that defense-adjacent settings can contain formal sequential decisions, but neither evaluates statistical stopping in a constructive-simulation or military operational-analysis workflow.

The five families provide partial ingredients for this unresolved problem. Precision or uncertainty contributions offer mechanisms for determining whether an estimate has reached a prescribed precision; discrete-decision and risk-or-loss contributions connect termination to actions or consequences; operational-state contributions include diagnostic and control gates; and resource-aware contributions make evaluations,

time, transmissions, or computational burden part of acquisition or termination. These elements cannot be combined by transferring guarantees between families or procedural branches. The absence of evaluations in defense constructive-simulation workflows leaves the open question: how to provide statistically justified, resource-aware, and interpretable stopping decisions for multiple operational outputs in defense constructive simulation.

C. PRACTITIONER GUIDANCE

Whereas Table 4 compares stopping mechanisms, Table 11 maps application constraints to candidate families, required conditions, and cautions. It provides conditional guidance, not a ranking or deployment recommendation; guarantees must be checked for the exact procedure and branch.

D. SURVEY LIMITATIONS

This survey has several limitations. First, the focal search covered Scopus and Web of Science through four targeted strategies and was restricted to English-language journal articles and conference papers published from 2015 through 2025. Records in languages other than English were therefore excluded. The survey also could not assess defense technical information absent from the searched bibliographic databases or unavailable for full-text assessment because of access controls. The Defense Technical Information Center, for example, holds material classified up to the Secret level and unclassified technical information subject to distribution controls [79]. These language and access boundaries may contribute to the observed defense gap, but their effect was not measured. These choices provided a reproducible route to operational stopping contributions, but they do not exhaust the broader literatures on sequential analysis, adaptive design, Markov chain Monte Carlo, Monte Carlo error control, or simulation output analysis. Publisher-specific platforms were not searched separately. Scopus and Web of Science nevertheless provided cross-publisher bibliographic coverage. The corpus included publications from 16 publishers or scientific societies, including Taylor & Francis, IEEE, Springer Nature, Elsevier, INFORMS, IET, and ACM. This cross-indexing reduced, but did not eliminate, the risk of platform-specific omissions because coverage remains selective by source, document type, and period. The extent of any such omission was not assessed. Full texts that could not be obtained and records outside the publisher boundary were tracked separately. The publisher boundary was an editorial rule for corpus construction and does not constitute a judgment about the scientific validity of every affected article or venue. The first author made the final decisions on study selection, contribution segmentation, family assignment, and evidence-state classification. The resulting corpus and classifications may therefore reflect single-assessor judgment.

Second, the contribution-level decomposition increased the resolution of the synthesis but required judgment. Distinct branches were separated when they differed materially in terminal objective, monitored quantity, stopping condition,

update mechanism, output, assumptions, or guarantees. Ambiguous relations were retained rather than normalized into artificial equivalence. Similarly, assigning contributions to families required identifying their principal terminal rationale. The resulting family profiles summarize the evidence documented for their constituent contributions and may conceal heterogeneity in mechanisms, assumptions, and evaluation settings. Family membership therefore does not transfer guarantees or evidence between contributions.

Third, the evidence appraisal describes what the reviewed studies reported. No formal study-level risk-of-bias instrument was applied. The five appraisal dimensions characterize reported validation and implementation evidence and were not designed or interpreted as a risk-of-bias assessment. The 95 contributions were not reimplemented on a common testbed, and the survey did not independently reproduce their theoretical guarantees, finite-sample results, computational measurements, or software behavior. Accordingly, a demonstrated rating indicates that the source study examined the dimension; it does not mean that the survey independently verified the result. The family-level profiles should not be interpreted as readiness scores, comparative rankings, deployment certifications, or universal recommendations.

Finally, independent validation and outside-domain evidence remain unresolved. The forward-citation assessment covered 53 source studies in Scopus and the remaining study through Web of Science, thereby covering all 54 source studies; candidate documents were screened from titles and abstracts. It did not verify full texts, institutional relationships, acknowledgments, funding, collaboration, the exact procedural variant implemented, or whether a later application represented a materially different domain. Relations without detected author overlap are therefore candidate signals rather than confirmed independent validation. Accordingly, the absence of defense constructive-simulation evaluations among the 95 focal contributions is an observation under the database, document-type, language, and access boundaries described above; it is not evidence that such work is absent from restricted collections, non-English literature, or elsewhere.

E. FUTURE DIRECTIONS

The research priorities concern external validation, changing data streams, and operational simulation. The survey does not resolve these gaps; it delineates research paths for addressing them, drawing on the evidence assessed and the starting points identified below.

1) Establishing Traceable External Evidence

The forward-citation assessment (Section II-G) provides an initial set of 31 documents for full-text examination, listed in *Forward-Citation Candidates for Full-Text Assessment* (Section IV-F). Assessment can begin with the 3 documents whose titles and abstracts indicate target-specific use or evaluation, followed by the remaining 28, which provide less explicit indications in their titles and abstracts. Full-text examina-

TABLE 11. Conditional guidance for relating practical constraints to the five stopping families.

When	Consider	Check	Main caution
A fixed precision is required for a mean or another scalar parameter.	Family 1 - Mean estimation / Other scalar-parameter estimation.	Target, tolerance, and sampling assumptions.	Meeting the precision target does not guarantee coverage.
Precision is required for a quantile or an extreme percentile.	Family 1 - Quantile estimation.	Quantile level, tail behavior, and dependence.	Extreme quantiles may require substantial effort and remain poorly estimated.
Several parameters must be estimated jointly.	Family 1 - Joint parameter estimation / Cross-cutting contributions.	Joint stopping condition and scope of the guarantee.	Requiring all marginal criteria to pass does not establish joint coverage.
Precision should scale with the magnitude or variability of the result.	Family 1 - Cross-cutting contributions; also the subfamily matching the target.	Whether the threshold is relative to the estimated value or to its variability.	Near-zero targets and unreliable scale estimates can impair stopping.
Several object estimates must meet a standard-error threshold.	Family 1 - Joint parameter estimation, for the collective object-estimation procedure represented in the corpus.	Per-object threshold, collective stopping condition, and comparison cap.	Some objects may reach the cap before satisfying the error threshold.
Sampling cost must be balanced against estimation error.	Family 3 - Risk or loss sufficiency.	Loss function and acquisition-cost model.	Optimality is specific to the model, assumptions, and limiting regime.
The result is a selection, classification, or detection.	Family 2 - Discrete-decision sufficiency.	Required decision guarantee and stopping boundary.	A cap-forced decision may not retain the regular guarantee.
A structure or output needs to stabilize.	Family 4 - Operational-state sufficiency.	Monitored property and stability threshold.	Apparent stability does not establish correctness or complete discovery.
Resource limits govern acquisition or stopping.	Family 5 - Resource-aware acquisition and stopping.	Resource measure, budget, and terminal condition.	Reduced resource use does not establish inferential sufficiency.

Note: Rows are not mutually exclusive. For contribution-level guidance, consult the “Consider when” column in the supplementary material (Section IV-F). Check the assumptions and limitations of the selected procedure before use.

tion would help determine whether these documents report evaluations of the stopping procedures and whether those evaluations provide independent or cross-domain evidence.

Future studies evaluating an existing stopping criterion should identify the operational variant implemented, document modifications to the original procedure, and specify the theoretical or empirical property under examination. Assessments of independence should consider the relationship between the originating and evaluating groups, while claims of transferability should describe the relevant differences in data-generating process, inferential objective, and operational context.

Such documentation would allow later studies to determine whether independently implemented procedures retain their reported properties and which assumptions matter outside the original setting. It would also permit subsequent reviews to distinguish methodological influence, reproducibility, independent validation, and cross-domain transfer as separate forms of evidence.

2) Characterizing Validity under Changing Data Streams

Further research should characterize the behavior of stopping criteria under heterogeneous and nonstationary data streams. This includes determining how changes in dependence and distribution affect stopping time, inferential validity, and resource consumption, as well as identifying the conditions under which an existing criterion can be recalibrated rather than reformulated.

A practical starting point is to compare stopping criteria that address the same inferential objective under controlled simulation conditions. An initial reference condition should satisfy the assumptions of the selected procedures. Subse-

quent experiments can introduce departures from those assumptions, first separately and then jointly, including changes in dependence, distributional shape, and variability. Extending these experiments to changes during data acquisition would help examine conditions that static comparisons do not capture. Across repeated independent simulation runs, the comparison should assess both the sampling effort and the statistical properties of the terminal estimates or decisions, rather than treating earlier stopping alone as evidence of improvement.

Particular attention is needed when structural changes and stopping sufficiency are assessed from the same accumulating observations. In this setting, uncertainty about the data-generating process must be considered when evaluating the stopping rule. Systematic evaluation across different dependence structures, degrees of heterogeneity, and regime transitions would help delimit where existing procedures remain reliable and where their guarantees cease to support the terminal decision.

3) Statistical Stopping in Operational Simulation

Defense constructive simulation provides a relevant setting for examining how statistical stopping can be incorporated into an established simulation workflow. Existing platforms already support repeated scenario execution, batch control, and the collection of multiple performance measures, but the available evidence does not establish how termination decisions should account simultaneously for inferential sufficiency, computational cost, and the operational interpretation of the outputs.

An initial application can focus on one operational scenario and one decision-relevant performance measure, selecting a

stopping criterion whose assumptions and terminal objective match that setting. The criterion can then be incorporated into batch execution and compared with a sufficiently extensive reference campaign. This comparison should examine whether the reduced simulation effort preserves the required accuracy or decision performance, while recording cases in which resource limits are reached before the stopping condition is satisfied. Subsequent evaluations can vary scenario conditions to determine how far the initial findings extend.

Further research should determine which statistical objectives are appropriate for terminating batches of constructive-simulation replications and how their validity is affected by scenario variation, dependence among outputs, and unequal replication costs. Evaluation in operationally representative workflows is also needed to quantify premature and delayed termination, the resources saved, and the consequences for the estimates or decisions produced. Such studies would establish whether statistical guarantees transfer to the conditions under which defense analyses are conducted.

F. SUPPLEMENTARY MATERIAL AVAILABILITY

The supplementary table *Operational Contributions and Evidence Profiles* provides the contribution-level assessment underlying the family and evidence profiles reported in this survey. Its “Consider when” column provides short target and application-condition phrases for locating relevant contributions. It covers the 95 operational contributions identified across the 54 primary studies. Supplementary materials are available at <https://github.com/a-nunes/stopping-criteria-simulation-survey>.

The supplementary table *Forward-Citation Candidates* lists the 31 external citing documents retained after the authorship-overlap assessment, with bibliographic metadata and links. It flags 3 documents as *Specific indication*, based on title-and-abstract indications of target-specific use or evaluation, and 28 as *Potentially relevant*. These records provide a starting point for follow-up assessment; they do not establish full-text relevance, independent validation, or cross-domain transfer.

V. CONCLUSION

Stopping is not a single notion of convergence: reaching a precision target, resolving a decision, accepting an operational state, and exhausting a budget justify different claims. A procedure can therefore finish without establishing the sufficiency its user requires. Selection must ask not only whether a rule saves effort, but what remains justified when acquisition ends. A narrow interval can miss its target, a stable structure can remain incomplete, and a budget-limited run can end with unresolved uncertainty. The implemented stopping condition and supporting evidence must justify the estimate, decision, or operational claim made at termination; family membership alone cannot do so.

For practitioners, this leads to a concrete sequence: define what must be established before stopping, use Table 11 to locate the relevant part of the taxonomy, and consult the

supplementary material (Section IV-F) to identify candidate contributions. Selection then depends on whether the procedure’s assumptions, guarantee scope, and reported limitations match the intended setting. Where that match is uncertain, the uncertainty should remain explicit rather than being concealed by the fact that a stopping threshold has been crossed.

The three directions in Section IV-E respond directly to the identified gaps: traceable external evidence and validity under changing data streams address the two cross-family gaps, while statistical stopping in operational simulation addresses the corpus-bounded application gap. The next step is to establish how reliably stopping decisions support their intended claims beyond the conditions in which they were developed. In defense constructive simulation, the focal corpus does not yet provide that application evidence. Evaluation in such workflows should examine not only the effort saved, but also the consequences of terminating too early or continuing unnecessarily. This would turn a plausible stopping mechanism into evidence for a justified decision about when enough simulation has been performed.

REFERENCES

- [1] A. M. Law, *Simulation Modeling and Analysis*, 5th ed., New York, NY, USA, 2015.
- [2] M. Jahangirian, T. Eldabi, A. Naseer, L. K. Stergioulas, and T. Young, “Simulation in manufacturing and business: A review,” *European Journal of Operational Research*, vol. 203, no. 1, pp. 1–13, 2010.
- [3] A. Negahban and J. S. Smith, “Simulation for manufacturing system design and operation: Literature review and analysis,” *Journal of Manufacturing Systems*, vol. 33, pp. 241–261, 2014.
- [4] T. Tang and T. A. Adcock, “Estimating space–time wave statistics using a sequential sampling method and Gaussian process regression,” *Applied Ocean Research*, vol. 122, p. 103127, 2022.
- [5] L. Gong and J. M. Flegal, “A Practical Sequential Stopping Rule for High-Dimensional Markov Chain Monte Carlo,” *Journal of Computational and Graphical Statistics*, vol. 25, no. 3, pp. 684–700, Jul. 2016.
- [6] Y. Zhou, H. Li, M. Chen, Z. Dai, and M. Zhu, “Accelerate tree ensemble learning based on adaptive sampling,” *Journal of Computational Methods in Sciences and Engineering*, vol. 20, no. 2, pp. 509–519, Jul. 2020.
- [7] J. P. A. Dantas, A. N. Costa, V. C. F. Gomes, A. R. Kuroswski, F. L. L. Medeiros, D. Geraldo, A. R. Samersla, S. R. Silva, A. F. M. Caetano, Y. D. Ferreira, A. O. Arantes, D. S. Santos, and M. R. C. Aquino, “ASA: A Simulation Environment for Evaluating Military Operational Scenarios,” in *Computational Science and Computational Intelligence*, ser. Communications in Computer and Information Science, H. R. Arabnia, L. Deligianidis, F. Shenavarmasouleh, S. Amirian, and F. Ghareh Mohammadi, Eds., Cham, 2025, vol. 2512, pp. 203–215, cSCI 2024, Las Vegas, NV, USA, December 11–13, 2024.
- [8] A. N. Costa, J. P. A. Dantas, E. Scutkins, F. L. L. Medeiros, and P. Ögren, “Simulation and Machine Learning in Beyond Visual Range Air Combat: A Survey,” *IEEE Access*, vol. 13, pp. 76 755–76 774, 2025.
- [9] M. Bicher, M. Wastian, D. Brunmeir, and N. Popper, “Review on Monte Carlo Simulation Stopping Rules: How Many Samples Are Really Enough?” *SNE Simulation Notes Europe*, vol. 32, no. 1, pp. 1–8, 2022.
- [10] D. I. Singham, “Selecting Stopping Rules for Confidence Interval Procedures,” *ACM Transactions on Modeling and Computer Simulation*, vol. 24, no. 3, pp. 1–18, May 2014.
- [11] A. M. Law, “Statistical Analysis of Simulation Output Data: The Practical State of the Art,” in *2020 Winter Simulation Conference (WSC)*, Orlando, FL, USA, Dec. 2020, pp. 1117–1127.
- [12] D. I. Singham, “Analysis of Sequential Stopping Rules for Simulation Experiments,” Ph.D. dissertation, University of California, Berkeley, Berkeley, CA, USA, 2010, engineering—Industrial Engineering and Operations Research. Committee chair: Lee Schruben. [Online]. Available: <https://escholarship.org/uc/item/3hb6p7bg>
- [13] N. Metropolis and S. Ulam, “The Monte Carlo Method,” *Journal of the American Statistical Association*, vol. 44, no. 247, pp. 335–341, Sep. 1949.

- [14] Y. S. Chow and H. Robbins, "On the Asymptotic Theory of Fixed-Width Sequential Confidence Intervals for the Mean," *The Annals of Mathematical Statistics*, vol. 36, no. 2, pp. 457–462, Apr. 1965.
- [15] G. S. Fishman, "Estimating Sample Size in Computing Simulation Experiments," *Management Science*, vol. 18, no. 1, pp. 21–38, Sep. 1971.
- [16] S. S. Lavenberg and C. H. Sauer, "Sequential Stopping Rules for the Regenerative Method of Simulation," *IBM Journal of Research and Development*, vol. 21, no. 6, pp. 545–558, Nov. 1977.
- [17] T. L. Moeller and P. D. Welch, "A spectral based technique for generating confidence intervals from simulation outputs," in *Proceedings of the 9th Conference on Winter Simulation*, Gaithersburg, MD, USA, Dec. 1977, pp. 177–184.
- [18] A. M. Law and W. D. Kelton, "Confidence Intervals for Steady-State Simulations II: A Survey of Sequential Procedures," *Management Science*, vol. 28, no. 5, pp. 550–562, May 1982.
- [19] —, "Confidence Intervals for Steady-State Simulations: I. A Survey of Fixed Sample Size Procedures," *Operations Research*, vol. 32, no. 6, pp. 1221–1239, Dec. 1984.
- [20] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hrobjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, L. A. McGuinness, L. A. Stewart, J. Thomas, A. C. Tricco, V. A. Welch, P. Whiting, and D. Moher, "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, p. n71, 2021.
- [21] M. L. Rethlefsen, S. Kirtley, S. Waffenschmidt, A. P. Ayala, D. Moher, M. J. Page, J. B. Koffel, and PRISMA-S Group, "PRISMA-S: An extension to the PRISMA statement for reporting literature searches in systematic reviews," *Systematic Reviews*, vol. 10, no. 1, p. 39, 2021.
- [22] H. Sie, M. D. Finkelman, J. Bartroff, and N. A. Thompson, "Stochastic curtailment in adaptive mastery testing: Improving the efficiency of confidence interval-based stopping rules," *Applied Psychological Measurement*, vol. 39, no. 4, pp. 278–292, 2015.
- [23] E. Park, "Comparison of sequential estimation in response-adaptive designs with and without covariate-adjustment," *Communications for Statistical Applications and Methods*, vol. 23, no. 4, pp. 287–296, 2016.
- [24] C. Alexopoulos, D. Goldsman, A. Mokashi, R. Nie, Q. Sun, K.-W. Tien, and J. R. Wilson, "Sequest: A sequential procedure for estimating steady-state quantiles," in *Proceedings of the 2014 Winter Simulation Conference*, 2015, pp. 662–673.
- [25] U. Bandyopadhyay and A. Biswas, "Sequential and two-stage fixed-width confidence interval estimation in response-adaptive designs," *Sequential Analysis*, vol. 34, no. 3, pp. 350–363, 2015.
- [26] I. R. Silva, D. Bhattacharjee, and Y. Zhuang, "Fixed-length interval estimation of population sizes: sequential adaptive monte carlo mark-recapture-mark sampling," *Computational and Applied Mathematics*, vol. 42, no. 4, p. 181, 2023.
- [27] C. Alexopoulos, D. Goldsman, A. Mokashi, K.-W. Tien, and J. R. Wilson, "Sequiem: Estimating extreme steady-state quantiles via the maximum transformation," in *2015 Winter Simulation Conference (WSC)*, 2016, pp. 562–574.
- [28] C. Alexopoulos, D. Goldsman, P. Tang, and J. R. Wilson, "Spsts: A sequential procedure for estimating the steady-state mean using standardized time series," *IIE Transactions*, vol. 48, no. 9, pp. 864–880, 2016.
- [29] C. Alexopoulos, D. Goldsman, A. C. Mokashi, and J. R. Wilson, "Automated estimation of extreme steady-state quantiles via the maximum transformation," *ACM Transactions on Modeling and Computer Simulation*, vol. 27, no. 4, pp. 1–29, 2017.
- [30] —, "Sequential estimation of steady-state quantiles: Lessons learned and future directions," in *2018 Winter Simulation Conference (WSC)*, 2018, pp. 1814–1825.
- [31] —, "Sequential estimation of steady-state quantiles: Some new developments in methods and software," in *2019 Winter Simulation Conference (WSC)*, 2019, pp. 3774–3785.
- [32] C. Alexopoulos, D. Goldsman, A. C. Mokashi, K.-W. Tien, and J. R. Wilson, "Sequest: A sequential procedure for estimating quantiles in steady-state simulations," *Operations Research*, vol. 67, no. 4, pp. 1162–1183, 2019.
- [33] U. Bandyopadhyay, S. Sarkar, and A. Biswas, "Fixed-width confidence interval of log odds ratio for joint binomial and inverse binomial sampling," *Sequential Analysis*, vol. 36, no. 3, pp. 345–354, 2017.
- [34] —, "Fixed-width confidence interval for treatment difference in multinomial sampling," *Sequential Analysis*, vol. 40, no. 2, pp. 198–208, 2021.
- [35] B. Chattopadhyay, "Accuracy in insurance billing error estimation using auxiliary information," *South African Statistical Journal*, vol. 54, no. 2, pp. 153–162, 2020.
- [36] H. Cho, "Two-stage procedure of fixed-width confidence intervals for the risk ratio," *Methodology and Computing in Applied Probability*, vol. 21, no. 3, pp. 721–733, 2019.
- [37] H. Cho and Z. Wang, "On fixed-width confidence limits for the risk ratio with sequential sampling," *American Journal of Mathematical and Management Sciences*, vol. 39, no. 2, pp. 166–181, 2020.
- [38] S. K. De and N. Mukhopadhyay, "Two-stage fixed-width and bounded-width confidence interval estimation methodologies for the common correlation in an equi-correlated multivariate normal distribution," *Sequential Analysis*, vol. 38, no. 2, pp. 214–258, 2019.
- [39] J. Dong and P. W. Glynn, "The Asymptotic Validity of Sequential Stopping Rules for Confidence Interval Construction Using Standardized Time Series," in *2019 Winter Simulation Conference (WSC)*, National Harbor, MD, USA, Dec. 2019, pp. 332–343.
- [40] D. A. Fitts, "Absolute precision confidence intervals for unstandardized mean differences using sequential stopping rules," *Behavior Research Methods*, vol. 55, no. 4, pp. 1839–1862, Jun. 2023.
- [41] J. M. Flegal and L. Gong, "Relative fixed-width stopping rules for markov chain monte carlo simulations," *Statistica Sinica*, vol. 25, no. 2, pp. 655–675, 2015.
- [42] S. Gürel, M. Şahin, İ. Uysal, A. İbileme, and T. Gündüz, "Adaptive selection algorithm and standard error termination rule in comparative judgement: An application for assessing writing skills," *Education and Science*, vol. 50, pp. 93–110, 2025.
- [43] J. Hu, "Improving hall's accelerated sequential procedure: Generalized multistage fixed-width confidence intervals for a normal mean," *Methodology and Computing in Applied Probability*, vol. 23, no. 3, pp. 823–835, 2021.
- [44] —, "A double-sequential sampling scheme," *Communications in Statistics - Theory and Methods*, vol. 51, no. 18, pp. 6319–6333, 2022.
- [45] K. Kelley, F. Bilson Darku, and B. Chattopadhyay, "Sequential accuracy in parameter estimation for population correlation coefficients," *Psychological Methods*, vol. 24, no. 4, pp. 492–515, 2019.
- [46] V. Maurya, A. Gill, and A. Goyal, "A new two-stage multiple comparison procedure for comparing several exponential populations with a control under heteroscedasticity," *Computational Statistics & Data Analysis*, vol. 108, pp. 1–11, 2017.
- [47] N. Mukhopadhyay and Y. Zhuang, "On fixed-accuracy and bounded accuracy confidence interval estimation problems in fisher's "nile" example," *Sequential Analysis*, vol. 35, no. 4, pp. 516–535, 2016.
- [48] N. Mukhopadhyay and J. Hu, "Two-stage estimation for a normal mean having a known lower bound of variance with final sample size defined via Gini's mean difference and mean absolute deviation," *Sequential Analysis*, vol. 37, no. 2, pp. 204–221, 2018.
- [49] N. Mukhopadhyay, "Asymptotic normality of sequential stopping times with applications: Confidence intervals for an exponential mean," *Calcutta Statistical Association Bulletin*, vol. 72, no. 1, pp. 17–34, 2020.
- [50] N. Mukhopadhyay and C. Zhang, "Eda on the asymptotic normality of the standardized sequential stopping times, part-ii: Distribution-free models," *Sequential Analysis*, vol. 39, no. 3, pp. 367–398, 2020.
- [51] N. Mukhopadhyay and S. K. Bishnoi, "On general asymptotically second-order efficient purely sequential fixed-width confidence interval (fwci) and minimum risk point estimation (mrpe) strategies for a normal mean and optimality," *METRION*, vol. 78, no. 3, pp. 383–409, 2020.
- [52] N. Mukhopadhyay and Z. Wang, "Purely sequential fwci and mrpe problems for the mean of a normal population by sampling in groups with illustrations using breast cancer data," *Sequential Analysis*, vol. 39, no. 2, pp. 176–213, 2020.
- [53] —, "Purely sequential estimation problems for the mean of a normal population by sampling in groups under permutations within each group and illustrations," *Sequential Analysis*, vol. 39, no. 4, pp. 484–519, 2020.
- [54] N. Mukhopadhyay and S. Venkatesan, "A new formulation of minimum risk fixed-width confidence interval (mrfwci) estimation problems for a normal mean with illustrations and simulations: Applications to air quality data," *Sequential Analysis*, vol. 41, no. 2, pp. 241–274, 2022.
- [55] N. Mukhopadhyay and A. Aloufi, "Second-order (s.o.) multi-stage fixed-width confidence interval (fwci) estimation strategies for comparing location parameters from two negative exponential (ne) populations: illustrations with cancer data," *Metrika*, vol. 87, no. 6, pp. 649–680, 2024.

[56] S. Sajjadippanah, E. Mahmoudi, and M. Zamani, "Two-stage procedure in a first-order autoregressive process and comparison with a purely sequential procedure," *Sequential Analysis*, vol. 40, no. 4, pp. 466–481, 2021.

[57] A. Steland, "On the accuracy of fixed sample and fixed width confidence intervals based on the vertically weighted average," *Journal of Statistical Theory and Practice*, vol. 11, no. 3, pp. 375–392, 2017.

[58] A. Steland and Y.-T. Chang, "High-confidence nonparametric fixed-width uncertainty intervals and applications to projected high-dimensional data and common mean estimation," *Sequential Analysis*, vol. 40, no. 1, pp. 97–124, 2021.

[59] T. Zoubeidi and M. Y. El-Bassiouni, "Fixed size confidence regions for the parameters of the mixed effects logistic regression model," *Electronic Journal of Applied Statistical Analysis*, vol. 12, no. 1, pp. 1–13, 2019.

[60] D. Batur and F. F. Choobineh, "Selecting the best alternative based on its quantile," *INFORMS Journal on Computing*, vol. 33, no. 2, pp. 441–456, 2021.

[61] G. Cui, L. Kong, W. Yi, B. Wang, and X. Yang, "Coupled sequential detection and estimation procedure for track-before-detect," *IET Radar, Sonar & Navigation*, vol. 10, no. 7, pp. 1168–1177, 2016.

[62] T. K. S. Solanky and J. Zhou, "A generalization of the partition problem," *Sequential Analysis*, vol. 34, no. 4, pp. 483–503, 2015.

[63] B. Karmakar and I. Mukhopadhyay, "Risk-efficient sequential estimation of multivariate random coefficient autoregressive process," *Sequential Analysis*, vol. 38, no. 1, pp. 26–45, 2019.

[64] L.-C. Hwang, "Asymptotic optimality of a robust two-stage procedure in multivariate bayes sequential estimation," *Sequential Analysis*, vol. 42, no. 1, pp. 43–55, 2023.

[65] A. Mukherjee, A. Tajer, P.-Y. Chen, and P. Das, "Active estimation from multimodal data," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5180–5184.

[66] Y. Zhuang and D. Bhattacharjee, "Minimum risk point estimation of the size of a finite population under mark-recapture strategy," *Sequential Analysis*, vol. 40, no. 2, pp. 243–258, 2021.

[67] A. S. Al-Sumaiti, M. H. Ahmed, S. Rivera, M. S. El Moursi, M. M. Salama, and T. Alsumaiti, "Stochastic pv model for power system planning applications," *IET Renewable Power Generation*, vol. 13, no. 16, pp. 3168–3179, 2019.

[68] L. A. Jiménez Rugama and L. Gilquin, "Reliable error estimation for sobol' indices," *Statistics and Computing*, vol. 28, no. 4, pp. 725–738, 2018.

[69] T. Schmalofski, M. Kroll, H. Dette, and R. Janisch, "Towards active learning: A stopping criterion for the sequential sampling of grain boundary degrees of freedom," *Materialia*, vol. 31, p. 101865, 2023.

[70] M. Tavana, D. Di Caprio, F. J. Santos-Arteaga, and A. O'Connor, "A novel entropy-based decision support framework for uncertainty resolution in the initial subjective evaluations of experts: The NATO enlargement problem," *Decision Support Systems*, vol. 74, pp. 135–149, 2015.

[71] H. Wu, W. Niu, S. Wang, and S. Yan, "An analysis method and a compensation strategy of motion accuracy for underwater glider considering uncertain current," *Ocean Engineering*, vol. 226, p. 108877, 2021.

[72] M. Semelhago, B. L. Nelson, E. Song, and A. Wächter, "Object-oriented implementation and parallelization of the rapid gaussian markov improvement algorithm," in *Proceedings of the 2022 Winter Simulation Conference*, 2022, pp. 3158–3169.

[73] C. Dang, P. Wei, J. Song, and M. Beer, "Estimation of failure probability function under imprecise probabilities by active learning-augmented probabilistic integration," *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, vol. 7, no. 4, p. 04021054, 2021.

[74] C. Quan, G. Joseph, and N. J. Myers, "An energy-efficient ordered transmission-based sequential estimation," in *2025 IEEE 101st Vehicular Technology Conference (VTC2025-Spring)*, 2025, pp. 1–6.

[75] Y. Zhao, Y. Han, Y. Liu, K. Xie, W. Li, and J. Yu, "Cross-entropy-based composite system reliability evaluation using subset simulation and minimum computational burden criterion," *IEEE Transactions on Power Systems*, vol. 36, no. 6, pp. 5198–5209, 2021.

[76] J. P. A. Dantas, S. R. Silva, V. C. F. Gomes, A. N. Costa, A. R. Samersla, D. Geraldo, M. R. O. A. Máximo, and T. Yoneyama, "AsaPy: A Python Library for Aerospace Simulation Analysis," in *Proceedings of the 38th ACM SIGSIM Conference on Principles of Advanced Discrete Simulation*, 2024, pp. 15–24.

[77] F. Lu, X. Hu, B. Zhao, X. Jiang, D. Liu, J. Lai, and Z. Wang, "Review of the research progress in combat simulation software," *Applied Sciences*, vol. 13, no. 9, p. 5571, 2023.

[78] A. T. Biggs, J. A. Hamilton, and R. R. Markwald, "Identifying appropriate scenario termination rules for squad-level simulations of warfighter lethality," *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology*, vol. 22, no. 4, pp. 521–528, 2025.

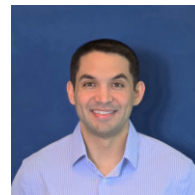
[79] Defense Technical Information Center, "Marking documents," <https://discover.dtic.mil/markings-documents/>, 2023, accessed: Sep. 20, 2026.



ARTUR LIGIERI NUNES received the B.Sc. degrees in Aeronautical Sciences and in Administration with emphasis in Public Administration from the Brazilian Air Force Academy (AFA), Pirassununga, Brazil, in 2013, and the M.B.A. degree in Leadership and Management from the Air Force University (UNIFA), Rio de Janeiro, Brazil, in 2024. He is currently pursuing the M.Sc. degree in Electronic and Computer Engineering at the Aeronautics Institute of Technology (ITA), Sao Jose dos Campos, Brazil. Nunes has served as a Brazilian Air Force career aviator since 2010 and currently holds the rank of Captain, with over 1,700 flight hours and experience in instruction, operational control, doctrine, and strategic and humanitarian missions. His research interests include simulation, statistical analysis of simulation output, and defense simulation.



GUILHERME MOREIRA is a Lieutenant Colonel in the Brazilian Air Force and a Professor at the Aeronautics Institute of Technology (ITA), Brazil. He holds B.Sc. degrees in Aeronautical Sciences, Administration, and Software Engineering, an M.Sc. in Aviation Safety from ISAE-SUPAERO/ENAC, France, and a Ph.D. in Aerospace Engineering from ITA. Moreira's experience includes operational certification at Airbus, certification and industrial processes for the F-X2/Gripen project, and avionics development management at the Institute of Aeronautics and Space. His current research focuses on the intersection of artificial intelligence, systems engineering, and defense simulation.



JOAO P. A. DANTAS received the B.Sc. degree in Mechanical-Aeronautical Engineering from the Aeronautics Institute of Technology (ITA), Brazil, in 2015. During his undergraduate studies, he participated in a year-long exchange program at Stony Brook University, USA. He obtained the M.Sc. degree in Electronic and Computer Engineering from ITA in 2019 and the Ph.D. degree in the same program in 2025. In 2022, he was a visiting researcher at the AirLab, part of the Robotics Institute at Carnegie Mellon University, USA. He currently works as a researcher for the Brazilian Air Force at the Institute for Advanced Studies. His research interests include machine learning, robotics, and simulation.



LUIZ ALBERTO VIEIRA DIAS received a degree in Electrical Engineering from the Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro, Brazil, in 1966, the M.Sc. degree in Space Science and Atmosphere from the National Institute for Space Research, Sao Jose dos Campos, Brazil, in 1968, the M.Sc. degree in Space Sciences from Rice University, Houston, TX, USA, in 1971, and the Ph.D. degree in Space Physics and Astronomy from Rice University in 1973. He is currently a Collaborating Professor with the Department of Software and Information Systems, Division of Computer Science, Aeronautics Institute of Technology (ITA), Sao Jose dos Campos, Brazil. His research interests include software engineering, applied computer engineering, numerical analysis, image processing, agile software development, and software testing.

...