

Article

Not peer-reviewed version

Legitimacy Stress-Testing AI-IoT Climate Adaptation: Contextual Validity and Incident Evaluation in Water Systems

[Victor Frimpong](#) *

Posted Date: 14 April 2026

doi: 10.20944/preprints202604.1005.v1

Keywords: AI governance; AI-IoT systems; water decision systems; legitimacy; climate adaptation; contextual validity; paper compliance risk; incident-risk evaluation; flood early warning



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Legitimacy Stress-Testing AI–IoT Climate Adaptation: Contextual Validity and Incident Evaluation in Water Systems

Victor Frimpong

Management Department, SBS Swiss Business School, Kloten, Switzerland
* Correspondence: v.frimpong@research.sbs.edu

Abstract

This paper argues that evaluating AI–IoT climate adaptation in water systems cannot rely solely on performance metrics; it requires legitimacy stress-testing grounded in contextual validity and incident-based assessment. While artificial intelligence (AI), machine learning, and Internet of Things (IoT) technologies are transforming water management—enhancing forecasting, monitoring, and decision-making for floods, droughts, and agricultural use—current evaluations remain largely model-centric, prioritising predictive accuracy over real-world viability. As a result, even technically robust systems can fail in practice, manifesting as missed events, false-alarm fatigue, delayed escalation, exclusion of vulnerable groups, and weak accountability—especially under climate variability and institutional constraints. The paper introduces a Legitimacy Stress-Test as a structured protocol for evaluating AI–IoT water systems as socio-technical infrastructures. Anchored in the Contextual Research Validity Index (CRVI), the framework comprises eight dimensions: data reliability, sensor performance, institutional readiness, governance of decision rights, equity, contestability, redress, and auditability. It links weaknesses across these dimensions to specific incident pathways, enabling proactive identification of governance risks and mitigation priorities. An illustrative flood early-warning case shows how strong predictive performance can fail to deliver resilience when contextual and governance conditions are misaligned. The proposed stress-test complements, rather than replaces, hydrological validation by clarifying when and why model performance breaks down. It offers a practical evaluation tool for agencies, donors, and regulators scaling AI–IoT climate adaptation systems.

Keywords: AI governance; AI–IoT systems; water decision systems; legitimacy; climate adaptation; contextual validity; paper compliance risk; incident-risk evaluation; flood early warning

1. Introduction

Climate change is intensifying hydrological extremes, resulting in more frequent and severe floods and droughts. These changes amplify flood vulnerability, extend drought duration, affect agricultural water requirements, and challenge water-supply reliability (Maharjan et al., 2025; Dahal et al., 2025). Increased rainfall variability, higher heat stress, and shifting seasonal patterns are heightening risks of flooding and water shortages, disrupting historically stable water patterns (Nazeri Tahroudi et al., 2025; IPCC, 2021). These challenges affect not just water resource management but also food security, disaster response, public health, and energy reliability. There is an urgent need for faster, more adaptive, and integrated decision-making in water management in the context of climate change.

The integration of artificial intelligence (AI), machine learning (ML), and the Internet of Things (IoT) is transforming climate-adaptation strategies in water resource management (Pasupuleti, 2025). IoT-enabled sensor networks provide continuous, real-time data on soil moisture, temperature, water levels, and water quality across river basins, reservoirs, irrigation systems, and urban drainage networks (Samadi, 2022; Alprol et al., 2024). These data streams are integrated with remote sensing and telemetry systems for effective watershed monitoring and situational awareness (Pasupuleti, 2025; Mukhamejanova et al., 2025).

ML-based predictive models enhance near-real-time forecasting, anomaly detection, and decision-making in water management (Kulkarni, 2025). This shift enables proactive water management, delivering timely flood alerts (Samadi, 2022), improved drought preparedness (Pasupuleti, 2025), and optimised irrigation scheduling (Alprol et al., 2024). It also facilitates coordination across water-energy-food (WEF) nexus systems. The combination of AI and IoT, known as the Artificial Intelligence of Things (AIoT), enhances WEF coordination by improving resource allocation and operational efficiency (Kulkarni, 2025; Wang & Abdelrahman, 2023). These advancements support greater climate resilience, sustainability, and risk-based planning (Pasupuleti, 2025).

1.1. The Implementation Gap: From Model Skill to Operational Outcomes

Despite advances in AI and IoT in water informatics, there remains a significant gap between model improvements and real-world outcomes. Studies highlight issues like scalability, data integrity, operational deployment, reliability, and effectiveness in practical settings, even when models perform well in controlled environments (Camacho-Leon, 2025; Dharmarathne et al., 2025). Much of the existing research focuses on model metrics such as RMSE, Nash–Sutcliffe efficiency, accuracy, and lead time (Samadi, 2022; Alprol et al., 2024). While these metrics are important for scientific validity and comparison, their interpretation can vary when applied to the same hydrological models (Samadi, 2022).

However, improved predictive accuracy does not guarantee reliable operational decisions, fair service delivery, or actual reductions in climate-related damages (Pasupuleti, 2025). This underscores the socio-technical nature of AI and IoT systems, which operate within institutional and political-economic contexts that shape whether forecasts lead to action and effective interventions (Pasupuleti, 2025; Mukhamejanova et al., 2025). To overcome challenges such as data integration, user engagement, stakeholder usability, and governance, interdisciplinary collaboration is essential for successful implementation (Mukhamejanova et al., 2025).

1.2. Governance, Legitimacy, and Paper Compliance in Climate-Sensitive Systems

This “bridge gap” is particularly significant in climate-sensitive systems, as failures can lead not only to inefficiency but also to damage. Flood early warning systems may demonstrate high predictive accuracy yet still face operational challenges, including delayed responses, ambiguous decision-making authority, fatigue from false alarms, insufficient outreach to at-risk communities, and weak accountability after events. Likewise, smart irrigation and drought decision-support systems may optimise water distribution theoretically but raise concerns about legitimacy in real-world applications, particularly when algorithmic results intersect with disputed resource allocation, ineffective grievance mechanisms, and unequal access to digital tools. In these scenarios, effectiveness hinges on governance and implementation conditions as much as on algorithmic performance.

A primary factor contributing to this discrepancy is that water AI research often regards governance, legitimacy, and institutional readiness as secondary issues rather than components to be evaluated in system performance (Pasupuleti, 2025). Governance is typically addressed in terms of “policy implications” or “stakeholder engagement,” but is seldom operationalised as a formal evaluation framework (Pasupuleti, 2025; Samadi, 2022). However, water systems are fundamentally socio-technical, integrating sensors and models with public institutions, legal frameworks,

emergency response protocols, budgetary constraints, and contested decision-making rights (Alprol et al., 2024). When these contextual factors are lacking, AI-IoT implementations may be technologically sophisticated yet operationally fragile, generating dashboards without approval, forecasts without escalation procedures, or data streams lacking maintenance provisions (Camacho-Leon, 2025). Challenges such as inadequate data management, poor explainability, and a lack of standardisation further limit effective implementation (Samadi, 2022). Under such circumstances, systems may appear compliant with oversight standards yet fail to enhance real-world resilience, creating a risk of “paper compliance”: governance documentation that lacks the operational capacity and legitimacy required for effective action.

1.3. Donor-Driven Deployment Pressures and Governance Deficits

The urgency of this problem is intensified by the rapid growth of climate-tech solutions. Water and climate adaptation increasingly rely on donor-funded digital infrastructure, public-private partnerships, and quick deployment timelines (Lazarus Jambadu et al., 2024). While this focus produces visible outputs such as models and dashboards, it often neglects critical governance elements, including maintenance budgets, decision rights, grievance procedures, and accountability frameworks (OECD, 2024; Lazarus Jambadu et al., 2024). Consequently, AI-IoT systems may fail to deliver expected benefits, erode trust, and hinder future adoption, particularly in contexts with weak institutional credibility.

1.4. Paper Contribution

This paper presents three key contributions. It introduces a Legitimacy Stress-Test, a protocol for evaluating AI-IoT water systems beyond model performance alone. Second, while Contextual Research Validity Index (Frimpong, 2026) provides the underlying theoretical logic of contextual validity, the Legitimacy Stress-Test operationalises this logic into an applied, domain-specific evaluation protocol for AI-IoT water systems. Third, it establishes an evaluation logic that links legitimacy deficits to predictable failure pathways.

1.5. Methodological Approach

We use a conceptual, abductive research design that integrates literature on AI governance, water informatics, and socio-technical systems to develop analytical frameworks. It develops the Legitimacy Stress-Test by combining key concepts—contextual validity, legitimacy, and contestability—and linking them to observable incident outcomes. Instead of empirical testing, the paper focuses on creating a theoretically sound and practical evaluation protocol, illustrated with examples to show its feasibility.

2. Conceptual Foundations: Contextual Validity, Legitimacy, and Incident-Oriented Evaluation

AI-IoT water decision systems are increasingly positioned as technical solutions to climate adaptation challenges (Dada et al., 2024; Petros Amanatidis et al., 2025). Yet their real-world performance depends not only on sensing accuracy and predictive skill but also on whether the system remains valid, legitimate, and actionable in deployment conditions.

2.1. Construct Clarification: Contextual Validity, Legitimacy, and Contestability

Contextual validity, legitimacy, and contestability are key concepts in the Legitimacy Stress-Test framework.

- Contextual validity assesses whether an AI-IoT system’s data, model assumptions, and design are appropriate for the specific local conditions, including environmental factors, infrastructure, and institutional settings.

- Legitimacy evaluates if the system's deployment and outcomes are considered acceptable and justified by stakeholders, including affected communities and oversight bodies.
- Contestability serves as a mechanism that maintains legitimacy amid uncertainty, allowing stakeholders to challenge and review system outputs through clear pathways for escalation and grievances.

Contextual validity concerns the trustworthiness of operations when employed in real-world settings, whereas *legitimacy* refers to the societal and institutional endorsement of the authority that supports the system's decisions. *Contestability* serves as a regulatory mechanism to avert problems that could lead to exclusion or governance failures.

The stress-test interprets contestability as a quantifiable procedural safeguard that preserves legitimacy when contextual validity is questionable or breaks down, rather than equating contestability directly with legitimacy.

2.2. AI-IoT Water Decision Systems as Socio-Technical Infrastructure

Water management and climate adaptation systems are fundamentally socio-technical (Shrimpton & Balta-Ozkan, 2024). They consist of physical infrastructure (such as rivers, reservoirs, canals, and pumping stations), digital elements (including sensors, telemetry, and cloud platforms), and institutional frameworks (like agencies, legal mandates, emergency protocols, budgets, and accountability mechanisms). AI and machine learning models are increasingly integrated into these systems as decision-support tools, providing forecasts, anomaly alerts, prioritisation outputs, or allocation suggestions.

Theoretically, integrating IoT sensing with AI analytics enhances situational awareness, improves early warning capabilities, and enables faster operational decision-making amid uncertainty (Guezouli et al., 2024; Tiggeloven et al., 2025). However, in practice, failures frequently occur even when the models are technically robust. For instance, a flood prediction may still lead to damage if there is ambiguity about escalation authority, responders lack trust in the alert, warnings do not reach at-risk populations, or sensor downtime compromises data reliability. Likewise, irrigation advice might be rejected if it is viewed as unjust, unclear, or inconsistent with local restrictions and water rights. These examples illustrate that AI-IoT implementations function not just as computational tools but also as governance frameworks.

Evaluations that focus solely on predictive metrics may overlook the institutional and social factors that determine whether forecasts lead to legitimate and effective actions. This underscores the need for a framework that treats contextual validity and legitimacy as essential performance dimensions rather than merely secondary implementation concerns.

2.3. Contextual Validity Under Climate Stress

The Contextual Research Validity Index (CRVI) assesses whether an analytical approach is credible in its operational setting, rather than presuming that methodological rigour alone ensures real-world validity (Frimpong, 2026). In AI-IoT water systems, contextual validity depends on multiple factors, including sensor reliability, data representativeness, maintenance capacity, institutional preparedness, and climate variability. In contrast to technical validity, which focuses on statistical generalisation and predictive performance, CRVI prioritises whether the system can maintain operational credibility amidst deployment limitations.

Climate change exacerbates this issue by altering historical rainfall patterns, runoff dynamics, and land-use interactions, thereby challenging the stationarity assumptions that often underpin hydrological validation. Even models that perform well in historical assessments may falter under regime shifts and extreme conditions. IoT sensing and real-time data streams can help mitigate this by facilitating ongoing monitoring, recalibration, and anomaly detection. Nonetheless, in resource-limited settings, contextual validity is often compromised by sensor outages, inadequate connectivity, procurement constraints, and data discontinuities, thereby diminishing the reliability

of model updates. As a result, contextual validity is not merely a matter of modelling but also tied to socio-institutional factors: it relies on whether institutions can sustain, interpret, and operationalise AI-IoT systems amid climate-related stresses (Pawson & Tilley, 1997).

2.4. Non-Stationarity and the Limits of Historical Validation

This challenge concerns non-stationarity in hydrological modelling, which is exacerbated by climate change. Changes in rainfall, runoff, drought duration, and the frequency of extreme events disrupt the assumptions of stationarity, complicating frequency analysis and risk assessment in water resource management (Twinomuhangi et al., 2025). Many basins show significant non-stationarity in runoff and sediment dynamics due to shifts in climate, which models based solely on historical data fail to capture (Wang et al., 2024). As a result, models may perform well during validation but struggle under different conditions, especially during extreme weather events that alter rainfall-runoff relationships (Twinomuhangi et al., 2025).

While IoT sensing and real-time data can help address non-stationarity by providing up-to-date observations, their effectiveness relies on data quality, coverage, and consistent operation. Integrating IoT-based monitoring with machine learning can improve real-time anomaly detection; however, these systems are hindered by network reliability and data integrity issues (Hashemi-Beni et al., 2024). Additionally, non-stationarity interacts with broader contextual factors: sensor failures may bias data and delay recalibration; inadequate governance may limit the use of predictions; and distrust or limited access may reduce effectiveness. Therefore, model performance is only one factor in real-world scenarios; socio-technical aspects also play a critical role (Hashemi-Beni et al., 2024).

2.5. Legitimacy and Contestability in Water Decision Systems

Legitimacy refers to the extent to which decision-making systems are perceived as acceptable by stakeholders within a governance framework (Haraway, 1988). It is also seen as the perceived appropriateness of authority and decision-making in line with social norms, values, and expectations (Suchman, 1995). In water management, legitimacy is particularly significant given the often contentious nature of resource allocation, varying levels of climate risk exposure, and power dynamics that shape trust and compliance. When AI is utilised for flood management, drought restrictions, or irrigation planning, its legitimacy hinges not only on its predictive accuracy but also on governance factors such as transparency, accountability, fairness, and the capacity of communities to contest its outputs (Cheong, 2024; Papagiannidis et al., 2025). In the absence of these elements, even systems with strong technical capabilities can undermine trust and become challenging to maintain, especially in environments where institutional credibility is already weak (Domínguez Hernández et al., 2025).

Contestability serves as a vital safeguard in algorithmic decision-making systems (Lyons et al., 2021). It denotes the capacity of individuals and communities to question results, request clarifications, seek corrections, and receive reparations. In water systems sensitive to climate change, contestability is not merely an ethical enhancement; it serves as a control mechanism that curtails the persistence of errors, reduces the risk of exclusion, and fosters learning. When agencies implement AI-IoT systems that lack clear accountability, grievance procedures, or visible post-incident correction mechanisms, stakeholders may perceive the technology as a means of exclusion or control rather than a tool for resilience (Walther, 2025). Consequently, the stress-test conceptualises legitimacy as a structured construct that includes (i) clarity in decision-making authority, (ii) transparency and auditability, (iii) equity and mechanisms to prevent exclusion, and (iv) contestability and redress options—each of which directly influences the responsible scaling of AI-IoT systems.

2.6. Paper Compliance Risk in AI Governance for Climate Adaptation

In this study, paper compliance risk refers to the gap between established AI governance documents and their practical application during implementation. This issue arises when policies, ethical guidelines, oversight bodies, or compliance paperwork are formally required but do not fit into decision-making workflows, such as escalation processes, override protocols, monitoring procedures, or post-incident analysis. Thus, the risk is defined not by the absence of governance materials but by their ineffectiveness in shaping operational practices (Pfeffer & Salancik, 1978), especially during stressful situations such as sensor malfunctions, model drift, or severe climate events.

Paper compliance risk is distinct from deliberate wrongdoing. It is more frequently a result of limitations in institutional capabilities, vague accountability, imported governance templates, or implementation pressures that favour visible outputs over actual operational effectiveness. In contexts of climate adaptation, where projects are often donor-funded, time-constrained, and involve multiple agencies, organisations may produce governance documentation that meets oversight requirements while leaving decision-making authority, maintenance duties, and learning processes underdeveloped.

In AI-IoT water management systems, paper compliance risks can manifest as dashboards without escalation authority, alerts without follow-up actions, models without financial support for maintenance, or incident analyses that do not lead to corrective measures. These situations result in systems that appear governed on paper yet remain organizationally vulnerable (Meyer & Rowan, 1977), thereby heightening the risk of recurring incidents and a breakdown of legitimacy. Aligning with broader discussions about gaps in governance practices highlighted in AI governance research (Frimpong, 2026), this paper considers paper compliance risk as a quantifiable vulnerability in deployment rather than merely a rhetorical statement.

2.7. Incident-Oriented Outcomes as the Missing Evaluation Layer

performance-centric assessment prioritises predictive accuracy; however, it often overlooks operational failures that only surface during actual deployment (Papagiannidis, 2025). In this paper, “incidents” denote real-world malfunctions and near misses in AI-IoT water decision systems, encompassing both technical failures (e.g., sensor outages, pipeline malfunctions, model drift, system downtime) and socio-institutional failures (e.g., delayed escalation, exclusion, breakdown of accountability, and loss of trust). Incidents are not limited to catastrophic events; they also encompass recurrent false alarms, unresolved complaints, and ongoing coverage gaps.

Incident-oriented evaluation transforms the inquiry from “How precise is the model?” to “What failures does the system generate in practice, and by what means?” This method aligns with operational risk management by focusing on failure pathways and control effectiveness rather than solely on model metrics. Practical outcomes can be monitored using indicators such as response time, false-alarm frequency, alert reach, override rates, completion of post-incident reviews, and closure of corrective actions.

The primary argument is that deficiencies in contextual validity and legitimacy are linked to predictable incident patterns. Unclear decision rights may lead to delayed escalation; weak contestability can allow errors to persist; limited auditability can diffuse accountability; and inadequate maintenance capacity can hasten sensor degradation and model drift. By linking governance and contextual shortcomings to incident outcomes, evaluators can identify failure mechanisms rather than merely reporting on performance metrics.

2.8. Summary of Conceptual Foundations

Prior discussion indicates that AI-IoT water decision-making systems should be assessed not only for their predictive accuracy but also for their contextual validity, legitimacy, and effectiveness in mitigating real-world incidents. In environments affected by climate stress, failures frequently

arise from the convergence of technological constraints (such as sensor failures, data inconsistencies, and non-stationarity) and governance shortcomings (including ambiguous decision-making authority, limited ability to contest decisions, insufficient audit capabilities, and risks associated with superficial compliance). This suggests the need for a systematic protocol to translate these socio-technical factors into quantifiable assessment standards. Consequently, the following section presents the Legitimacy Stress-Test, a scoring framework akin to the CRVI that features established dimensions and clear connections to incident pathways, intended for both pre-event readiness evaluation and post-incident analysis.

To ensure clarity regarding the paper’s foundations and to maintain consistency across the dimensions of the stress-test and evaluation logic, Table 1 presents the essential constructs, their definitions, and sample operational indicators used in the study.

Table 1. Key Constructs and Operational Indicators.

Construct	Working definition (This paper)	Why it matters in AI-IoT water systems	Illustrative operational indicators
AI-IoT water decision system	An integrated water/climate management system using sensing (IoT/remote sensing), analytics (AI/ML), and decision processes (Pasupuleti, 2025; Gamit et al., 2024)	These systems serve as governance infrastructure, not merely forecasting tools.	Sensor uptime, model update frequency, and alert-to-action workflow clarity
	The degree to which system outputs stay relevant, reliable, and practical within local institutional, infrastructural, and climate contexts (Frimpong, 2026).	Determines if model skill results in operational usefulness.	Data completeness, maintenance, climate non-stationarity stress-testing.
Legitimacy	Perceived acceptability of the system's decisions and processes is influenced by fairness, accountability, transparency, and contestability (Suchman, 1995).	Conditions: adoption, compliance, and trust in warnings/allocations	Stakeholder acceptance, grievance use, and responsibility clarity.
Contestability	Affected parties can challenge outputs, request explanations, seek corrections, and access redress (Lyons et al., 2021).	Prevents error persistence and minimises the risk of exclusion.	Appeal channel existence, time-to-resolution, and correction rate.

Paper compliance risk	The risk of strong governance in documentation but weak in practice (<i>Papagiannidis, 2025</i>).	Produces systems that “look governed” but fail in incidents.	Governance artefacts lack SOP integration, have unclear escalation rights, and are missing post-event learning.
Incident-oriented outcomes	Real-world failures and near misses during system use (<i>Papagiannidis, 2025</i>).	Captures failure modes that model metrics miss.	Missed events, false alarms, delayed escalation, exclusion, and accountability issues.
Decision rights clarity	The extent to which authority for action, override, and escalation is explicitly assigned (<i>Pasupuleti, 2025; Gamit et al., 2024</i>).	Determines whether forecasts lead to timely action	Who can issue alerts, override frequency, and escalation latency?
Auditability	Traceability of system outputs, decisions, and actions for accountability and learning (<i>Papagiannidis, 2025</i>).	Enables post-incident analysis and assignment of responsibility.	Logs, decision records, and completed post-incident reviews.
Equity/exclusion risk	The probability that system design or implementation disadvantages certain groups (<i>Haraway, 1988</i>).	Water and climate decisions often spread risk and scarcity.	Alert on group reach, access gaps, and inclusion metrics.

3. The Legitimacy Stress-Test: Design and Scoring Framework

The Legitimacy Stress-Test protocol is designed to evaluate AI-IoT water decision systems under climate stress. This tool addresses a prevalent challenge in climate-tech implementations: systems may exhibit technical proficiency yet encounter operational difficulties and legitimacy issues due to contextual misalignment, insufficient decision-making authority, or inadequate governance design. The Legitimacy Stress-Test assesses contextual validity and legitimacy and links them to predictable incident pathways.

3.1. Design Principles and Intended Use

- The Legitimacy Stress-Test is designed for both initial and follow-up assessments.
- Ex ante application (pre-deployment / scaling):
The tool serves as an audit to assess whether an AI-IoT system will remain valid and legitimate in its intended context and whether the governance design is sufficiently mature for safe scaling.
 - Ex post application (post-deployment / operational monitoring):

The tool serves as a diagnostic instrument to elucidate the causes of incidents that occur despite robust predictive performance and to identify the governance and contextual enhancements required to prevent recurrence.

Designed as a CRVI-style scoring system, it translates abstract notions of governance and legitimacy into explicit, evidence-based criteria. This approach enables evaluators, regulators, agencies, donors, and implementation partners to do consistent, verifiable assessments.

3.2. Unit of Analysis

The focus is on the deployed AI-IoT decision system rather than the model itself. This includes:

1. Sensing layer: IoT sensors, telemetry, remote sensing inputs, data conduits.
2. Analytics layer: machine learning models, hybrid models, digital twins, uncertainty estimation.
3. Decision layer: alert thresholds, allocation criteria, escalation procedures, override privileges.
4. Institutional layer: governance, ownership, mandates, budgets, personnel, responsibility.
5. Social layer: contestability, grievance procedures, equitable implications, trust dynamics.

This broadened unit of analysis acknowledges that incidents arise from the interactions among these levels rather than solely from model inaccuracies.

a. Dimensions of Contextual Validity and Legitimacy

The Legitimacy Stress-Test has eight key dimensions that identify how AI-IoT water systems can fail despite good technical design. Each dimension is defined for scoring based on evidence such as documents, operational logs, maintenance plans, stakeholder protocols, and audit trails.

1. Climate robustness and non-stationarity readiness
2. Data representativeness and measurement integrity
3. Sensor reliability and maintenance sustainability
4. Institutional capacity and operational readiness
5. Decision rights, escalation pathways, and override governance
6. Equity and exclusion risk controls
7. Contestability, grievance, and redress mechanisms
8. Transparency, auditability, and post-incident learning

These dimensions reflect the system's legitimacy and its effectiveness in reducing incidents.

3.4. Scoring Logic and Evidence Requirements

Each dimension is rated on a 0–3 scale, with higher scores indicating greater legitimacy and a lower risk of incidents.

- 0 = Absent/high-risk (governance missing or purely symbolic)
- 1 = Partial/fragile (exists but incomplete, informal, or under-resourced)
- 2 = Adequate/functional (operationally integrated, evidence exists)
- 3 = Robust/resilient (tested under stress, monitored, and continuously improved)

The scoring is not meant to label systems as “good” or “bad.” Instead, it serves as a diagnostic tool for structured comparison and highlights key risk areas that need attention.

3.5. Interpreting Stress-Test Outputs

The tool generates three practical outputs:

1. Legitimacy risk profile: a dimension-by-dimension scorecard identifying weak points.
2. Incident pathway prediction: mapping low-scoring dimensions to likely operational failures.
3. Mitigation priorities: a ranked set of governance and implementation improvements required before scaling.

This output design ensures the tool is not merely descriptive but also actionable. Expanding on these concepts, **Table 2** operationalises the stress-test by translating each dimension into anchored scoring criteria (0–3), enabling consistent cross-system evaluation. Scores are assigned based on

verifiable evidence such as SOPs, governance charters, monitoring reports, incident logs, and after-action reviews.

Table 2. The Legitimacy Stress-Test (CRVI-Style) – Dimensions and Anchored Scoring Criteria. Dimensions D1–D3 highlight the importance of technical contextual validity, whereas D4–D8 focus on operational legitimacy and governance resilience, which encompasses auditability and learning from post-incident events. Contestability is regarded as a fundamental safeguard that spans various areas, primarily reflected in D7 and reinforced by D5 and D8. For dimensions without evidence, cap the score at 1.

Dimension	0 = Absent / High-risk	1 = Partial / Fragile	2 = Adequate / Functional	3 = Robust / Resilient
D1. Climate robustness and non-stationarity readiness	Assumes stationarity; no stress-testing for climate shifts; model considered stable.	Acknowledges non-stationarity but relies on ad hoc adjustments; limited stress tests.	Includes stress-testing with multiple climate/extreme scenarios and documented update triggers.	Continuous drift monitoring, regime-change protocol, and model recalibration integrated into operations.
D2. Data representativeness and measurement integrity	Training and operational data are biased, incomplete, or unknown, with no quality controls.	Basic data cleaning with partial coverage and unclear bias/exclusion risks.	Outlined data QA/QC, assessed representativeness, and documented limitations.	Systematic monitoring, bias controls, adaptive sampling, and validation.
D3. Sensor reliability and maintenance sustainability	Sensors are deployed without maintenance plans, resulting in frequent downtime and inadequate lifecycle budgeting.	A maintenance plan exists but is unfunded; spare parts procurement is weak; calibration is irregular.	Funded maintenance plan; uptime monitored; calibration schedule implemented.	Resilient operations include redundancy, quick repairs, tamper detection, and secured lifecycle funding.
D4. Institutional capacity and operational readiness	No trained operators, unclear ownership, no SOP integration, and reliance on external consultants.	Some training and SOPs exist, but capacity remains fragile and reliant on key individuals.	Clear ownership, staffing, SOPs, internal competence, and defined roles.	Institutionalised capacity: training pipeline, turnover resilience, cross-agency coordination, performance oversight.
D5. Decision rights, escalation, and	Alerts exist but lack an authority pathway. Unclear	Some escalation rules; overrides possible but	Decision-rights charter, escalation thresholds, and	Tested in drills; override logs are reviewed;

override governance	decision rights and overrides are informal.	undocumented; responsibility is ambiguous	override rules are documented and operationally integrated.	escalation performance is monitored and continuously improved.
D6. Equity and exclusion risk controls	No analysis of exclusion, the system presumes equal access, and vulnerable groups remain invisible.	Equity mentioned but not implemented; limited outreach; no monitoring.	Vulnerable groups are identified; alert reach and coverage are tracked; mitigation plans are in place.	Equity built in: multilingual channels, offline redundancy, targeted dissemination, and ongoing inclusion monitoring. High-functioning contestability:
D7. Contestability, grievance and redress mechanisms	No channel for challenge or correction; complaints ignored.	Informal feedback is possible but inconsistent, with slow or unclear resolution.	Formal grievance process with a set resolution time and documented correction procedures.	rapid correction, feedback integrated into model/governance updates, transparency to complainants.
D8. Transparency, auditability and post-incident learning	No logs, decision traceability, or post-event learning.	Partial logs; learning depends on individuals; reviews are irregular.	Audit trail exists, post-incident reviews done, corrective actions tracked.	ML system: routine reviews, accountability mapping, continuous improvement.

3.6. Evidence Requirements and Inter-Rater Scoring Protocol

The rubric emphasises transparency in scoring by relying on documented evidence rather than evaluator impressions. To enhance scoring consistency, the protocol aims to bolster inter-rater reliability by using evidence-based anchors rather than relying on evaluators' impressions. Each score must be supported by verifiable artefacts (such as SOPs, governance charters, monitoring reports, incident logs, after-action reviews, and audit trails). Where such evidence is lacking or merely declaratory, the score for that dimension will be limited to 1.

For formal evaluations, the paper recommends that two evaluators conduct independent scoring in parallel, followed by a reconciliation step to discuss and resolve disagreements based on the rubric criteria. If disagreements continue, assessors are advised to note a brief contested scoring comment that outlines the evidence in question and the reasoning behind the differing scores. This method ensures transparency, enhances replicability, and enables future empirical assessment of interrater reliability.

3.7. Why the Tool Adds Value Beyond Accuracy

The Legitimacy Stress-Test complements technical model evaluation by identifying conditions under which improvements in accuracy do not translate into improved resilience outcomes. It also helps detect compliance risks. In climate-sensitive water systems, which often face non-stationarity, institutional fragility, and contested decision rights, this tool promotes responsible scaling and enhances the credibility of AI-IoT deployments.

The subsequent section establishes a connection between these dimensions and a structured incident taxonomy. It also proposes measurable operational indicators that empower evaluators to transition from model-centric metrics to a more incident-oriented approach to performance assessment.

4. Incident-Oriented Evaluation: Linking Contextual Legitimacy Deficits to Operational Failure Pathways

4.1. Defining Incidents in AI-IoT Water Decision Systems

This section presents an incident-oriented evaluation framework designed to enhance model-centric validation by addressing real-world system failures.

4.2. Defining Incidents in AI-IoT Water Decision Systems

An incident-oriented evaluation layer complements model-centric performance metrics (Schwartz et al., 2025). While metrics such as root mean square error (RMSE), Nash–Sutcliffe efficiency (NSE), and lead time are important for validating AI-IoT models (Samadi, 2022; Alprol et al., 2024), they do not adequately assess real-world harm reduction. In climate-sensitive water systems, operational outcomes depend on contextual constraints, decision authority, trust, and institutional readiness. This means that technically sound models can still lead to operational failures.

Incident-oriented evaluation shifts focus from model outputs to actual system behaviour in the real world. It examines deployment failures, their causes, and the contextual factors that predict them. This approach aligns with risk management and operational resilience, in which safety and effectiveness are evaluated based on incident pathways, near misses, and control performance, rather than solely on accuracy.

4.3. Defining Incidents in AI-IoT Water Decision Systems

In this study, an incident is an operational failure or near miss caused by an AI-IoT decision system. It encompasses both:

- Technical incidents: sensor inoperability, data pipeline malfunction, model degradation, system unavailability, cyber disruption.
- Socio-institutional incidents: protracted escalation, alarm fatigue, marginalisation of vulnerable populations, failure of accountability, disputed allocation results, and deterioration of confidence.

This definition is intentionally broad. In climate adaptation, major failures often stem from a confluence of technical difficulties and governance deficiencies, where governance deficits exacerbate technical inadequacies, leading to legitimacy crises.

4.4. A Practical Incident Taxonomy

The study employs six categories of occurrences based on common failure patterns in early warning systems, drought management, and irrigation decision-making (Papagiannidis, 2025).

1. Incidents of missed events: absence of alert, delayed alert, or alert falling below the actionable threshold.
2. Incidents of false alarm fatigue: excessive alerts reduce attention and compliance.
3. Delayed escalation incidents: alarms are triggered, but action is deferred due to ambiguous authority or lack of cooperation.

4. Exclusion incidents: marginalised groups are either not notified of notifications or unable to respond due to accessibility barriers.
5. Incidents of accountability breakdown: ambiguity in responsibility, absence of audit records, or deficiencies in post-event learning.
6. Operational drift incidents: the system is utilised beyond its validated parameters (e.g., altered basin conditions, new climatic regimes, revised operational aims) without appropriate governance adjustments.

These categories are intended to be applicable across domains, including flood early warning, drought allocation, reservoir operations, irrigation scheduling, and urban water management.

4.5. Why Incident-Oriented Outcomes Are the Missing Link

Accuracy-focused evaluation often presumes that better forecasts lead directly to improved outcomes, but this assumption is increasingly questioned in real-world AI deployment scenarios (Schwartz et al., 2025). Relying solely on accuracy benchmarks often overlooks significant operational failures and potential harms that can occur during deployment (Papagiannidis, 2025; Dharmarathne et al., 2025). For example, this assumption is invalid in circumstances where:

- data are unreliable or biased,
- sensor networks degrade,
- institutions lack readiness,
- decision rights are unclear,
- affected groups do not trust or cannot contest outputs, or
- accountability mechanisms do not exist.

In these circumstances, the capacity to forecast results diverges from operational efficacy. Thus, incident-oriented results provide the fundamental evaluative criterion: they illustrate system behaviour under duress and evaluate the extent to which the governance structure facilitates action, equity, and learning.

4.6. Mapping Legitimacy Stress-Test Dimensions to Incident Pathways

We suggest that deficiencies in legitimacy and contextual relevance might predict emerging trends. **Table 3** links stress-test dimensions to failure mechanisms, incident outcomes, and measurable indicators, making governance risks empirically traceable. For example:

- Poor maintenance sustainability prolongs sensor downtime, causing missed events and operational drift.
- Weak decision rights and escalation governance cause more delayed escalation incidents despite accurate alerts.
- Weak grievance mechanisms heighten exclusion and mistrust, leading to greater non-compliance.
- Weak auditability undermines accountability, harms legitimacy, and blocks learning.

This paradigm enables assessors to identify not only the occurrence of incidents but also the governance deficiencies that increased their probability. The framework also facilitates proactive forecasting: systems that receive low ratings in specific areas are prone to generating specific types of incidents unless remedial actions are implemented.

Table 3. Dimension-to-Incident Pathway Map: Mechanisms and Operational Metrics.

Stress-test dimension	Failure (How breaks)	mechanism legitimacy	Likely outcomes	incident	Minimal indicators (ex post)	measurable
-----------------------	----------------------	----------------------	-----------------	----------	------------------------------	------------

D1. Climate robustness and non-stationarity readiness	Model becomes invalid during regime shifts; drift isn't monitored; thresholds become uncalibrated.	Missed events, operational drift, repeated forecast failures during extremes.	Drift detection, recalibration lag, forecast errors during extremes, threshold update intervals.
D2. Data representativeness and measurement integrity	Bias or missingness causes systematic error, leading to model underperformance in certain zones/groups.	Missed events, exclusion incidents, uneven risk detection	Errors by sub-region, missingness rate, bias audit results, and coverage map gaps.
D3. Sensor reliability and maintenance sustainability	Sensor issues like downtime, tampering, or calibration failures degrade data quality and timeliness.	Missed events, false alarms, system downtime	Sensor uptime %, calibration adherence, repair time, redundancy, and data latency.
D4. Institutional capacity and operational readiness	System is not integrated into SOPs, relying on individuals; staff cannot interpret or act.	Delayed escalation, accountability issues, and inconsistent responses.	Response latency, staff training, SOP compliance, and handover continuity.
D5. Decision rights, escalation, and override governance	Alerts fail to trigger action; authority is vague; overrides happen informally; blame is diffused.	Delayed escalation, missed events, and accountability breakdown.	Time-to-escalation, override logs and reason, drill performance, decision chain audit.
D6. Equity and exclusion risk controls	Vulnerable groups not reached; digital divide; warnings not actionable; unequal allocation impacts.	Exclusion incidents, trust erosion, and non-compliance	Alert reach by group; channel coverage (SMS, radio, community); surveys on action feasibility; complaint concentration by group.
D7. Contestability, grievance and redress mechanisms	Errors cannot be challenged; complaints unresolved; no correction loop	Exclusion incidents, persistent errors, and a legitimacy crisis.	Existence of grievance channel, resolution time, correction rate, and repeat complaint rate.
D8. Transparency, auditability and post-incident learning	Decisions lack traceability and accountability, with recurring failures and	Accountability issues, recurring incidents, and erosion of trust.	Log completeness, post-incident review rate, corrective action closure, and recurrence of audit findings.

superficial
governance.

4.7. *What Evaluators Should Always Collect*

The research presents a minimum assessment set that can be collected without advanced infrastructure, facilitating practical implementation. The indicators are classified into four categories:

- i. *Operational performance*
 - response latency (alert → action)
 - escalation time (alert → responsible authority engaged)
 - system availability/uptime
- ii. *Signal quality in use*
 - false alarm rate in operational context
 - missed event rate (including late alerts)
 - alert fatigue indicators (e.g., ignored alerts over time)
- iii. *Inclusion and legitimacy*
 - alert reach by vulnerability group
 - complaint rate and complaint concentration
 - contestation resolution time
- iv. *Accountability and learning*
 - audit trail completeness
 - post-incident review completion
 - corrective action closure rate

These indicators are designed to complement hydrological evaluation metrics rather than supplant them, ensuring that both operational performance and legitimacy are evaluated alongside model accuracy.

4.8. *Transitioning from Legitimacy Deficits to Predictable Incident Risk*

The section illustrates how gaps in legitimacy and contextual validity lead to predictable incident pathways. This means that evaluators can utilise the Legitimacy Stress-Test not only to outline governance conditions but also to predict which failures are likely to occur unless appropriate mitigations are implemented.

5. Illustrative Case: Applying the Legitimacy Stress-Test to an AI-IoT Flood Early Warning System

This case-based walkthrough demonstrates how to use the Legitimacy Stress-Test in an AI-IoT water decision framework. The aim is to show the tool’s practical application: evaluators can assess contextual legitimacy, identify incident paths, and suggest governance improvements. This approach supports the integration of advanced modeling with practical use and offers a protocol for evaluators.

5.1. *Case Rationale: Why Flood Early Warning is an Appropriate Illustrative Setting*

Flood early warning systems provide an exemplary framework for implementing the Legitimacy Stress-Test for three primary reasons. Flood situations necessitate critical, time-sensitive decisions, in which governance structures and the authority to escalate can profoundly influence outcomes. Secondly, contemporary flood early warning systems frequently employ IoT sensors (including rain gauges and river-level sensors), remote sensing technology, and AI prediction models, demonstrating the synergy between AI and IoT. Ultimately, an effective flood response requires collaboration among diverse stakeholders, including meteorological agencies, disaster

management organisations, local government, and community leaders, and the establishment of a governance framework that prioritises legitimacy, contestability, and accountability.

5.2. Illustrative System Description

The system is a regional flood early-warning system implemented across a river basin, supporting both urban and rural communities. The architecture of the system includes:

- Sensing layer: river stage sensors and rainfall gauges are connected through cellular telemetry, with additional data provided by satellite precipitation products.
- Analytics layer: A hybrid forecasting system that integrates machine learning-driven rainfall-runoff predictions with rule-based thresholds to trigger alerts.
- Decision layer: Automated alert generation via SMS and dashboard notifications, including recommended escalation if predicted levels surpass predefined thresholds.
- Institutional layer: A primary water agency overseeing system operations, coordinating with disaster management authorities and local governments.
- Social layer: community-level recipients, such as high-risk populations residing in floodplains.

The system is considered to have performed well in controlled validation assessments and exhibits significant predictive ability for river-level exceedances. However, its performance in real-world conditions varies, with observed issues including false alarms, delayed responses, and inconsistent alert distribution.

5.3. Applying the Stress-Test: Evidence and Scoring Approach

A thorough evaluation requires multiple sources of evidence, including system logs, maintenance records, standard operating procedures, training materials, escalation protocols, stakeholder engagement documentation, and audit trails. The scores presented are illustrative, reflecting typical early-stage deployment conditions in the implementation literature, and do not constitute a validated assessment of any specific system. The goal is to show how the scoring protocol creates a clear legitimacy risk profile, identifies potential incident pathways, and aids in prioritising governance improvements.

5.4. Results: Legitimacy Risk Profile and Predicted Incident Pathways

Table 4 presents the scoring profile for initial deployments. It indicates that although technological capability exists, governance and contextual preparedness are deficient. The resulting risk profile can indicate likely specific incident outcomes, including delayed escalation, false-alarm fatigue, and exclusion of vulnerable groups.

Table 4. Illustrative Stress-Test Scorecard for an AI-IoT Flood Early Warning System.

Dimension	Illustrative score (0–3)	Key evidence assumptions	Primary incident risk
D1. Climate robustness and non-stationarity readiness	1	Model validated on historical data; limited extreme stress-testing.	Missed events under extreme regimes.
D2. Data representativeness and measurement integrity	2	Adequate basin coverage; some gaps in upstream rural zones.	Uneven detection; sub-region blind spots.
D3. Sensor reliability and maintenance sustainability	1	Downtime occurs, calibration is inconsistent,	Missed events; false alarms due to sensor noise

			and spare parts procurement is slow.	
D4.	Institutional capacity and operational readiness	1	System operated by a small team; SOPs exist but are weakly embedded.	Delayed escalation; inconsistent response
D5.	Decision rights, escalation, and override governance	0	Alert issued, but no clear authority to declare evacuation or mobilise resources.	Delayed escalation; accountability breakdown
D6.	Equity and exclusion risk controls	1	SMS alerts reach owners, but offline channels are weak, and language barriers exist.	Exclusion incidents; unequal warning reach
D7.	Contestability, grievance and redress mechanisms	0	No formal mechanism to challenge false alarms or missed alerts.	Persistent errors; trust erosion
D8.	Transparency, auditability and post-incident learning	1	Partial logs, inconsistent post-event review, and untracked corrective actions.	Repeated incidents; blame diffusion

5.5. Interpretation: Why “Strong Models” Still Produce Weak Outcomes

The system demonstrates strong predictive performance in controlled settings, but stress tests indicate a high risk of operational failure. Low scores in decision rights and escalation governance (D5), contestability (D7), and institutional readiness (D4) indicate that, although forecasts are made, they often do not prompt timely action, allowing errors to persist without proper correction. This indicates a compliance risk: although governance may be documented, actual authority and accountability for effective responses are lacking.

The scorecard also highlights a negative feedback loop between technical and institutional weaknesses. For instance, sensor downtime and calibration issues (D3) raise the likelihood of false alarms and missed events. Without contestability mechanisms (D7), communities and frontline responders lack formal avenues to report inaccuracies or request corrections, leading to persistent errors and a loss of trust. Over time, this situation increases non-compliance and diminishes the effectiveness of early warnings, despite the underlying model being statistically sound.

5.6. Translating the Scorecard into Mitigation Priorities

An important aspect of the Legitimacy Stress-Test is that it translates a risk profile at the dimension level into specific governance priorities. According to the example scorecard, three actions could significantly reduce the likelihood of incidents without requiring immediate modifications to the forecasting model.

Decision-rights charter and escalation protocol (D5):

Create a decision-rights map that clearly outlines who can escalate alerts, recommend evacuations, mobilise resources, and authorise overrides. Include specific escalation thresholds, handover procedures, and accountability for any delays in action (Frimpong et al., 2025).

Grievance and redress mechanism (D7):

Establish a formal process for communities and frontline authorities to report false alarms, missed alerts, and access issues. This process should include clear timelines for resolution and corrective action, thereby creating a feedback loop that minimises errors and maintains trust.

Maintenance and reliability plan with lifecycle budgeting (D3):

Secure funded maintenance schedules for calibration, spare parts, and repairs, with redundancy in high-risk areas. This minimises downtime and enhances the stability of sensing and alerting systems.

Further enhancements include regular monitoring of drift and climate stress evaluations (D1), along with equity-centred alert distribution (D6) through multilingual communication, offline backup methods (such as radio and local community contacts), and audits of coverage among at-risk populations.

Worked Example: How the Stress-Test can Indicate likely Incident Outcomes

Scenario: Heavy rainfall upstream causes rapid runoff and rising river levels. The AI-IoT system sends an alert 6 hours before the threshold is exceeded.

Observed outcome: Evacuations began late despite an early alert, resulting in preventable losses in several low-income areas.

Explanation of stress testing:

- *D5 (Decision rights = 0):* An alert is issued, but no agency has the authority to escalate to evacuation orders or mobilise emergency resources. While a forecast is available, action is delayed due to negotiations over who is responsible.
- *D4 (Operational readiness = 1):* Local authorities lack training and proper SOP integration, causing confusion in interpretation and response.
- *D6 (Equity controls = 1):* SMS alerts often fail to reach many households, and there is no offline distribution available.
- *D8 (Auditability = 1):* Logs are incomplete after the incident, and there is no corrective action register, leading to recurring failures in future events.

Incident classification: Delayed escalation, exclusion, and breakdown in accountability.

Mitigation: Implementing a decision-rights charter, offline alert redundancy, and post-incident review protocols can help reduce recurrence without modifying the existing model.

5.7. Summary of What the Walkthrough Demonstrates

The Legitimacy Stress-Test identifies hidden operational weaknesses that standard model evaluations may miss. Even strong predictive performance can mask issues in decision rights, institutional readiness, contestability, equity controls, and auditability, leading to problems like delayed escalation, exclusion, false-alarm fatigue, and accountability failures. By creating an evidence-based scorecard for legitimacy and contextual readiness, the stress-test helps diagnose and prioritise governance improvements, ensuring the responsible scaling of AI-IoT water decision systems amid climate stress.

6. Implications for Procurement, Governance, and Climate Adaptation Evaluation

The Legitimacy Stress-Test influences the procurement, assessment, governance, and expansion of AI-IoT climate-adaptation systems, particularly in water decision-making, where failures can rapidly lead to adverse consequences. This section elucidates how agencies, funders, regulators, and implementers might utilise the stress test to mitigate compliance risks and bolster resilience against incidents.

6.1. Reframing "Performance" in AI-IoT Water Systems

The key takeaway from this paper is that "performance" is multi-layered. While model accuracy and predictive skill are important, they alone don't indicate resilience. In practical situations, performance relies on whether:

- sensing and data pipelines remain reliable over time,
- institutions can interpret and act on outputs,
- decision rights are clearly assigned,
- alerts reach vulnerable populations, and
- post-incident learning is institutionalised.

Focusing assessments on technical measures may yield systems that appear effective in evaluations but falter in critical contexts. This underscores the necessity for evaluative standards that incorporate contextual legitimacy and actual incident outcomes as essential performance metrics.

6.2. A CRVI-Style “Legitimacy Annexe” for Procurement and Scaling

The stress test can serve as a procurement and scaling tool by mandating the inclusion of a Legitimacy Annexe in AI-IoT project documentation. The annex should include:

- the dimension-by-dimension scorecard (Table 2 structure),
- evidence supporting each score,
- identified incident pathways, and
- mitigation commitments required prior to scaling.

This method offers three practical benefits. It establishes a systematic approach to uncover hidden implementation risks early. Secondly, it minimises dependence on informal judgment or ad hoc stakeholder accounts. Lastly, it enables auditing of legitimacy and contextual validity, facilitating comparability across proposals and implementations.

In real-world applications, agencies and donors could mandate minimum standards for scaling. For instance, scaling could depend on achieving at least “adequate” ratings (≥ 2) on decision rights (D5), contestability (D7), and auditability (D8), due to their strong connection with delayed escalation and accountability failures.

6.3. Decision Rights as a Governance Primitive in Climate-Tech

The stress-test identifies decision rights and escalation governance (D5) as key factors influencing incident outcomes, thereby informing governance design for flood warning, drought allocation, and irrigation systems.

AI-IoT systems often assume that information is associated with action, but an effective response requires clear authority assignments: who can escalate issues, override decisions, mobilise resources, and take accountability for failures. Without this clarity, the system becomes merely informational rather than operational. As a result, early warnings may be issued, but responses are often delayed due to unclear authority and responsibilities.

A Decision Rights Charter is a necessary tool for deployment readiness. This charter should clearly outline:

- escalation thresholds and responsible authorities,
- override rules and conditions,
- accountability mapping across agencies, and
- minimum response time expectations (service-level commitments).

This instrument is particularly important in multi-agency water governance, where overlapping responsibilities and fragmented mandates often occur.

6.4. Incident Registers and Post-Event Learning as Operational Governance

AI-IoT systems should be managed through incident registers and structured learning cycles, rather than relying solely on compliance checklists. This approach treats these systems as operational risks, similar to safety-critical infrastructure.

A basic governance requirement is to maintain an Incident and Near-Miss Register to track events:

- missed events,

- false alarms and alert fatigue,
- delayed escalation cases,
- exclusion failures, and
- accountability breakdown events.

Each incident entry should include the timeline, contributing factors related to the stress-test dimensions, corrective actions taken, and closure status. This helps with learning and minimises recurrence.

Agencies should implement After-Action Reviews (AARs) for major events:

- what the model predicted,
- what alerts were issued,
- who acted (and when),
- where escalation failed, and
- What governance or capacity changes are required?

Without a robust learning infrastructure, systems often repeat the same mistakes, thereby increasing mistrust and hindering adoption.

6.5. Contestability and Redress as Resilience Infrastructure

Contestability is crucial in water and climate systems, not only as an ethical requirement but also as a component of resilience infrastructure. Errors in flood warnings, drought restrictions, or irrigation allocations are inevitable. The key issue is the ability to detect, challenge, correct, and learn from these errors.

To address this, AI and IoT systems should incorporate a Grievance and Redress Mechanism as a standard operational component. This mechanism should include:

- clear reporting channels (digital and offline),
- defined time-to-resolution targets,
- correction procedures for erroneous alerts or exclusions, and
- feedback integration into the model and governance updates.

This mechanism reduces instances of exclusion and enhances trust, especially in situations where resource distribution is disputed and institutional legitimacy is weak.

6.6. Equity Controls: From Inclusion Rhetoric to Measurable Reach

Many climate-tech initiatives mention inclusion but often lack measurable equity controls. The stress test suggests that equity should be treated as an operational performance metric rather than merely a rhetorical commitment.

A practical tool for this is an Equity Reach Audit, which requires evaluators to measure:

- alert reach by vulnerability group,
- channel redundancy (SMS, radio, community focal points),
- language accessibility, and
- action feasibility (whether recipients can act on warnings).

Audits can be conducted with minimal data collection using community surveys, dissemination logs, and geographic coverage mapping. This is especially important for flood warning systems, as high-risk populations often live in informal settlements with limited digital access.

6.7. Donor and Public–Private Partnership Governance: Reducing “Pilotitis”

Many AI–IoT water systems are often funded by donors or through public–private partnerships. While these arrangements can lead to quick pilots and visible results, they frequently neglect long-term maintenance and capacity building. To reduce unnecessary repetition among pilots, the stress-test requires implementers to demonstrate their readiness for sustained success.

Donors and implementers must prioritise the following essential conditions for scaling:

- funded maintenance and lifecycle budgets (D3),

- internal operational capacity (D4),
- decision rights and escalation protocols (D5), and
- auditability and learning cycles (D8).

In the absence of these criteria, systems might perform well during demonstration stages but fail when exposed to real-world climate extremes, leading to a loss of trust and wasted investment.

6.8. Regulatory and Evaluation Implications

The stress-test affects regulators and standard-setting bodies. As AI-IoT systems play a larger role in climate adaptation and public safety, evaluation frameworks may need to include more than just technical validation. Regulators might consider incorporating the stress-test into their assessments:

- climate-tech certification processes,
- disaster preparedness audits,
- public digital infrastructure governance requirements, or
- procurement standards for AI-enabled water systems.

The tool enables proportional governance by targeting specific factors that predict operational incidents and legitimacy breakdown, rather than imposing generic compliance requirements.

6.9. Strengthening the Bridge Between Modelling and Implementation

The Legitimacy Stress-Test can be integrated into actual governance procedures. By adding a Legitimacy Annexe to procurement, a Decision Rights Charter for managing escalations, an Incident Register for knowledge acquisition, and tools to measure equity and contestability, organisations can lower the risk of mere paper compliance and enhance real-world resilience.

7. Conclusion

AI-IoT water decision systems are being used as climate-adaptation solutions, but their effectiveness goes beyond just accurate predictions. Many technically sound systems encounter operational issues, including missed events, false alarms, and accountability breakdowns, due to factors such as sensor fragility and weak governance. These challenges highlight a significant evaluation gap: standard assessments often fail to determine if AI forecasts lead to actionable decisions under stress.

This paper introduces a Legitimacy Stress-Test, a scoring framework designed to evaluate these AI-IoT systems as socio-technical infrastructures, rather than isolated models. The framework includes eight dimensions: climate robustness, data integrity, sensor reliability, operational readiness, governance of decision rights, equity controls, contestability and redress, and auditability with lessons learned post-incident. By connecting weaknesses in these dimensions to potential failure pathways, this tool enables a shift from abstract governance principles to practical mechanisms, helping measure compliance risks as operational vulnerabilities.

The stress-test can create a legitimacy risk profile, identify potential failure pathways, and prioritise governance interventions to reduce incidents without requiring changes to the forecasting model. This provides a practical evaluation tool for agencies and regulators to responsibly scale AI-IoT climate systems, especially in areas with uneven institutional capacity and legitimacy.

The Legitimacy Stress-Test is designed to complement existing hydrological validation metrics. Its goal is not to predict incident probabilities with precision but to facilitate transparent auditing of the conditions that affect model performance. Future research should empirically validate this framework and apply it to incident data in flood, drought, and irrigation contexts. The protocol provides a practical framework for improving procurement, monitoring, and learning in climate-tech deployments, thereby enhancing the link between advanced modelling and real-world legitimacy.

References

1. Alprol, A. E., Mansour, A. T., Ibrahim, E. M., & Ashour, M. (2024). Artificial Intelligence Technologies Revolutionising Wastewater Treatment: Current Trends and Future Prospective. *Water*, 16(2), 314. <https://doi.org/10.3390/w16020314>
2. Amanatidis, P., Lyratzis, E., Angelopoulos, V., Kouloumpis, E., Skaperdas, E., Bassiliades, N., Vlahavas, I., Maris, F., Emmanouloudis, D., & Karampatzakis, D. (2026). Intelligent Water Management Through Edge-Enabled IoT, AI, and Big Data Technologies. *IoT*, 7(1), 5. <https://doi.org/10.3390/iot7010005>
3. Camacho-Leon, S. (2025). Emerging trends in IoT for aquatic systems: a systematic literature review. *Frontiers in Water*, 7. <https://doi.org/10.3389/frwa.2025.1699240>
4. Cheong, B. C. (2024). Transparency and accountability in AI systems: safeguarding wellbeing in the age of algorithmic decision-making. *Frontiers in Human Dynamics*, 6. <https://doi.org/10.3389/fhumd.2024.1421273>
5. Dada, M. A., Majemite, M. T., Obaigbena, A., Daraojimba, O. H., Oliha, J. S., Nwokediegwu, Z. Q. S., Dada, M. A., Majemite, M. T., Obaigbena, A., Daraojimba, O. H., Oliha, J. S., & Nwokediegwu, Z. Q. S. (2024). Review of smart water management: IoT and AI in water and wastewater treatment. *World Journal of Advanced Research and Reviews*, 21(1), 1373–1382. <https://doi.org/10.30574/wjarr.2024.21.1.0171>
6. Dahal, D., Bhattarai, N., Silwal, A., Shrestha, S., Shrestha, B., Poudel, B., & Kalra, A. (2025). A Review on Climate Change Impacts on Freshwater Systems and Ecosystem Resilience. *Water*, 17(21), 3052. <https://doi.org/10.3390/w17213052>
7. Dharmarathne, G., Abekoon, A. M. S. R., Bogahawaththa, M., Alawatugoda, J., & Meddage, D. P. P. (2025). A review of machine learning and internet-of-things on the water quality assessment: Methods, applications and future trends. *Results in Engineering*, 26, 105182. <https://doi.org/10.1016/j.rineng.2025.105182>
8. Domínguez Hernández, A., Perini, A. M., Hadjiloizou, S., Borda, A., Mahomed, S., & Leslie, D. (2025). Towards a sociotechnical ecology of artificial intelligence: power, accountability, and governance in a global context. *AI and Ethics*, 6(1). <https://doi.org/10.1007/s43681-025-00902-6>
9. Frimpong, V. (2026). The Challenge of Context-Free Validity: Introducing the Contextual Research Validity Index Framework for Situated Legitimacy under Socioeconomic Challenges. *SocioEconomic Challenges*, 10(1), 42-49. [https://doi.org/10.61093/sec.10\(1\).42-49.2026](https://doi.org/10.61093/sec.10(1).42-49.2026)
10. Frimpong, V., & Mamuti, A. (2026). Adaptive Methodologies and Bricolage Research Design in Africa's Digital Asymmetry. *International Journal of Applied Research in Business and Management*, 7(1). <https://doi.org/10.51137/wrp.ijarbm.429>
11. Haraway, D. (1988). Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies*, 14(3), 575–599. <https://doi.org/10.2307/3178066>
12. Hashemi-Beni, L., Puthenparampil, M., & Jamali, A. (2024). A low-cost IoT-based deep learning method of water gauge measurement for flood monitoring. *Geomatics, Natural Hazards and Risk*, 15(1). <https://doi.org/10.1080/19475705.2024.2364777>
13. IPCC. (2021). Weather and Climate Extreme Events in a Changing Climate. In *Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 1513–1766). Chapter, Cambridge: Cambridge University Press. United Kingdom and New York, NY, USA
14. Lahcene Guezouli, Lyamine Guezouli, Baha, M., & Abir Bentahrour. (2024). IoT and AI for Real-time Water Monitoring and Leak Detection. *Journal of Renewable Energies*, 27(2). <https://doi.org/10.54966/jreen.v27i2.1210>
15. Lazarus Jambadu, Jochen Monstadt, & Pilo, F. (2024). The politics of tied aid: Technology transfer and the maintenance and repair of water infrastructure. *World Development*, 175, 106476–106476. <https://doi.org/10.1016/j.worlddev.2023.106476>
16. Lyons, H., Velloso, E., & Miller, T. (2021). Conceptualising Contestability. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 1–25. <https://doi.org/10.1145/3449180>
17. Maharjan, S., Li, W., Bolten, J. D., & El-Askary, H. (2025). The future intensification of hydrological extremes and whiplashes in the contiguous United States increases community vulnerability. *Communications Earth & Environment*, 6(1). <https://doi.org/10.1038/s43247-025-02672-9>

18. Meyer, J., & Rowan, B. (1977). Institutionalised Organisations: Formal Structure as Myth and Ceremony. *American Journal of Sociology*, 83, 340–363. <http://dx.doi.org/10.1086/226550>
19. Mukhamejanova, A., Kazhimkanuly, D., Utepov, Y., & Aniskin, A. (2025). Modern technologies for monitoring waterlogged areas and their impact on infrastructure. *Bulletin of L N Gumilyov Eurasian National University Technical Science and Technology Series*, 150(1), 226–248. <https://doi.org/10.32523/2616-7263-2025-150-1-226-248>
20. OECD (2024), *Infrastructure for a Climate-Resilient Future*, OECD Publishing, Paris, <https://doi.org/10.1787/a74a45b0-en>.
21. Papagiannidis, E., Mikalef, P., & Conboy, K. (2025). Responsible artificial intelligence governance: A review and research framework. *The Journal of Strategic Information Systems*, 34(2). <https://doi.org/10.1016/j.jsis.2024.101885>
22. Pasupuleti, M. K. (2025). AI for Climate Resilience: Predictive Analytics for Global Risk Reduction and Sustainable Development. *International Journal of Academic and Industrial Research Innovations (ijairi)*, 54–66. <https://doi.org/10.62311/nesx/rp05116>
23. Pawson, R., & Tilley, N. (1997). An introduction to scientific realist evaluation. In E. Chelmsky & W. R. Shadish (Eds.), *Evaluation for the 21st century: A handbook* (pp. 405–418). Sage Publications, Inc. <https://doi.org/10.4135/9781483348896.n29>
24. Pfeffer, J. & Salancik, G. (1978). *The External Control of Organisations: A Resource Dependence Perspective*. Harper & Row, New York.
25. Samadi, S. (2022). The convergence of AI, IoT, and big data for advancing flood analytics research. *Frontiers in Water*, 4. <https://doi.org/10.3389/frwa.2022.786040>
26. Schwartz, R., Chowdhury, R., Kundu, A., Frase, H., Fadaee, M., David, T., Waters, G., Taik, A., Briggs, M., Hall, P., Jain, S., Yee, K., Thomas, S., Bhandari, S., Duncan, P., Thompson, A., Carlyle, M., Lu, Q., Holmes, M., & Skeadas, T. (2025). Reality Check: A New Evaluation Ecosystem Is Necessary to Understand AI's Real World Effects. *ArXiv.org*. <https://arxiv.org/abs/2505.18893>
27. Shashank Dattatray **Kulkarni** (2025). Artificial Intelligence and Machine Learning Applications in Precision Agriculture: Enhancing Crop Yield Prediction, Disease Detection, and Resource Optimisation. *International Journal of Applied Mathematics*, 38(11s), 1874-1887. <https://doi.org/10.12732/ijam.v38i11s.1300>
28. Shrimpton, E. A., & Balta-Ozkan, N. (2024). A Systematic Review of Socio-Technical Systems in the Water–Energy–Food Nexus: Building a Framework for Infrastructure Justice. *Sustainability*, 16(14), 5962. <https://doi.org/10.3390/su16145962>
29. Suchman, M. C. (1995). Managing Legitimacy: Strategic and Institutional Approaches. *The Academy of Management Review*, 20(3), 571–610. <https://doi.org/10.2307/258788>
30. Tahroudi, M. N. (2025). Comprehensive global assessment of precipitation trend and pattern variability considering their distribution dynamics. *Scientific Reports*, 15(1). <https://doi.org/10.1038/s41598-025-06050-5>
31. Tiggeloven, T., Pfeiffer, S., Matanó, A., van den Homberg, M., Thalheimer, L., Reichstein, M., & Torresan, S. (2025). The Role of Artificial Intelligence for Early Warning Systems: Status, Applicability, Guardrails and Ways Forward. *IScience*, 113689. <https://doi.org/10.1016/j.isci.2025.113689>
32. Twinomuhangi, M. B., Bamutaze, Y., Kabenge, I., Wanyama, J., Kizza, M., Gabiri, G., & Egli, P. E. (2025). Analysis of stationary and non-stationary hydrological extremes under a changing environment: A systematic review. *HydroResearch*, 8, 332–350. <https://doi.org/10.1016/j.hydres.2024.12.007>
33. Walther, C. C. (2025, January 28). How AI Exclusion Impacts Humankind. *Knowledge at Wharton*; [Knowledge@Wharton](https://knowledge.wharton.upenn.edu/article/how-ai-exclusion-impacts-humankind/). <https://knowledge.wharton.upenn.edu/article/how-ai-exclusion-impacts-humankind/>
34. Wang, Q. & Abdelrahman, W. (2023). High-Precision AI-Enabled Flood Prediction Integrating Local Sensor Data and 3rd Party Weather Forecast. *Sensors*, 23(6), 3065. <https://doi.org/10.3390/s23063065>
35. Frimpong, V.; Tawk, C.; Mamuti, A. (2025). Human–AI Handovers: A Dynamic Authority Reversal (DAR) Framework for Trust Calibration and Transitional Accountability. In *Proceedings of the IX. International Applied Social Sciences (C-IASOS-2025) Congress*, Rome, Italy, 13–15 October 2025.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.