Principles for Responsible Application of Artificial Intelligence to Support Climate Adaptation Metrics, Monitoring, Evaluation, Reporting, and Learning

Authors: Dr. Samraj Sahay and Karl Schultz
International Platform on Adaptation Metrics (IPAM)

Publication information and document context

This front matter records the publication status, intended use and institutional context for the Principles.

Document	Principles for Responsible Application of Artificial Intelligence to Support Climate Adaptation Metrics, Monitoring, Evaluation, Reporting, and Learning
Issued by	International Platform on Adaptation Metrics (IPAM)
Authors	Dr. Samraj Sahay and Karl Schultz
Version	Launch version
Date	9 July 2026
Intended audience	Practitioners, evaluators, technical programme managers, consultants, funders, researchers, policymakers and organizations applying or interpreting AI-supported analytical outputs in adaptation metrics and MERL.
Document status	Professional guidance. The Principles support transparent and responsible use of AI but do not certify that AI-generated outputs are correct and do not create binding obligations.

Purpose: The Principles are intended to help adaptation metrics and MERL professionals show that AI was used transparently, under human oversight, with attention to data quality, bias, contextual validity and documented limitations.

Suggested citation

Sahay, S. and Schultz, K. (2026). Principles for Responsible Application of Artificial Intelligence to Support Climate Adaptation Metrics, Monitoring, Evaluation, Reporting, and Learning. International Platform on Adaptation Metrics (IPAM). Launch version, 9 July 2026.

How to use this document

- Use the Principles when selecting, configuring, verifying or interpreting ready-made AI/ML services in adaptation metrics and MERL.
- Treat the Essential Principles as a minimum due diligence threshold for AI-assisted professional work.
- Use the risk glossary and implementation guidance to decide when non-technical users should escalate to technical, legal or security leads.
- Adapt the Principles to local context, institutional capacity and the stakes of the decision being supported.

Acknowledgements

IPAM gratefully acknowledges the IPAM AI Principles Task Group for its deliberation, drafting contributions and review of these Principles. IPAM also thanks all IPAM and non-IPAM members who contributed feedback, served as discussants and completed the IPAM survey. Their inputs guided the creation and improvement of the Principles and helped strengthen their usability for practitioners and others applying or interpreting AI-supported analytical outputs in adaptation metrics and MERL.

Preparation process

The Principles were prepared through an IPAM task group process and refined through discussion, review and survey feedback. This process was used to clarify the intended users and deployers of the Principles, identify trade-offs in their application, and strengthen treatment of credibility, data quality, bias, human oversight and contextual validation.

About the International Platform on Adaptation Metrics (IPAM)

The International Platform on Adaptation Metrics (IPAM) was launched in 2020 following adaptation metrics convenings initiated by the Moroccan Presidency of the Conference of Parties to the UN Framework Convention on Climate Change (COP22). IPAM serves as an international reference platform for adaptation metrics.

IPAM seeks to co-develop metrics and tools in response to emerging adaptation needs and to create synergies among its members. Its work addresses sector-oriented priorities, such as agriculture, cities and water, as well as cross-sectoral metrics harmonisation and the identification and application of appropriate techniques and tools for metrics.

Website: www.adaptationmetrics.org

> **Context:** This document extends IPAM's focus on adaptation metrics into the responsible use of AI for metrics, monitoring, evaluation, reporting and learning. It is designed as a practical starting point for responsible AI use, not as a final compliance regime.

1. Introduction

In recent years, artificial intelligence (AI) has emerged as a transformative technology with a potential to bring about impactful changes in achieving adaptation to climate risks. With the need for assessing the progress on the Global Goal on Adaptation (GGA) under Article 7 of the Paris Agreement and the targets set by the United Arab Emirates-Framework for Global Climate Resilience (UAE-FGCR), the potential importance of integrating AI for adaptation measurement has immensely increased. The use of AI for identifying adaptation metrics and the development of monitoring, evaluation, reporting and learning (MERL) can have beneficial effects on achieving these targets. This, however, comes with several challenges and risks such as ethical issues of bias embedded in AI-generated outputs, leading to inequality and non-inclusiveness. Further concerns relate to data security, lack of transparency, accuracy and validity of the outcomes, and misinformation that could have adverse impacts for measuring adaptation progress. This document identifies key challenges and establishes pragmatic Principles for safe, effective, and credible use of AI for the adaptation metrics and MERL. It uses the available evidence and inputs from an *expert survey* (see Appendix A2) with an intention to make the Principles more usable and oriented toward practitioners and others applying or interpreting AI-supported analytical outputs in adaptation work.

For evaluators, technical program managers, consultants, and experts developing adaptation metrics frameworks, the IPAM AI Principles provide a practical basis for demonstrating minimum due diligence in the use of AI. They do not certify that AI-generated outputs are correct. Rather, they help practitioners show that AI was used transparently, under human oversight, with attention to data quality, bias, contextual validity, and documented limitations. In this sense, the Principles strengthen the credibility, defensibility, and trustworthiness of AI-assisted professional work.

The document has been divided into the following sections: Section 2 provides the statement on *scope for the* '*Key Principles for Responsible Use of AI'* for adaptation metrics and MERL that IPAM wants to come up with for the adaptation community. Section 3 lays down the list of major challenges and risks associated with the use of AI for adaptation metrics. Section 4 details out the boundary or scope for the principles governing how users should employ AI services in Adaptation-Metrics and MERL. Finally, Section 5 lists the *Key Principles for Responsible Use of AI'.*

2. Scope of the Principles

IPAM's mission is to provide experts involved in adaptation metrics, and the related practices of monitoring, evaluation, reporting and learning (MERL) with a forum to discuss and co-produce in the subject areas, to formulate and enhance good practices. As AI application for metrics and MERL increases alongside rapid technological evolution, both opportunities and risks grow. This creates a compelling need for guidance on appropriate AI applications.

As such, the scope of this Statement of Principles is focused and limited to support equitable, transparent, and scientifically robust application of AI. It does not advocate for or against specific uses of AI, while recognizing the emerging nature of AI and climate adaptation metrics, and thus the need for a dynamic and evolving set of principles.

- This Statement of Principles is intended to guide '*deployers*' of the AI and end '*users*' of the AI who benefit from the deployers' use of the Principles for AI in adaptation metrics and MERL.
    - The '*deployers'* include Researchers and technical advisers, Project teams using AI, Evaluators or MERL Specialist and Organizational leaders and managers.
    - The end '*users'* include Practitioners and implementers, Policymakers and planners, Evaluators or MERL Specialist, and Researchers and technical advisers
- Principles are intended to guide the selection, configuration, verification and interpretation of available AI services. They are not intended to guide AI technology or application developers, even for AI applications supporting adaptation metrics and MERL.
- The Principles are valid for the broad set of tools under the 'artificial intelligence' umbrella, including Large Language Models, Machine Learning, and big-data analytics.
- The Principles may be used for adaptation metrics and MERL activities, and the document may be referenced to provide clarity and transparency on how AI is applied.
- Principles should be adapted to local contexts and individual and institutional capacities.
- The Statement of Principles is a starting point offered by IPAM. As technologies, laws and policies change, and experiences inside and out of climate adaptation research and practice grow, this document should be revisited.

Disclaimer:

Use of the Statement of Principles is intended as guidance rather than enforceable standards, and it does not create binding obligations on IPAM and the authors of these principles for actions taken or decisions made based on this Statement

3. AI for fostering adaptation – the adaptation metrics perspective

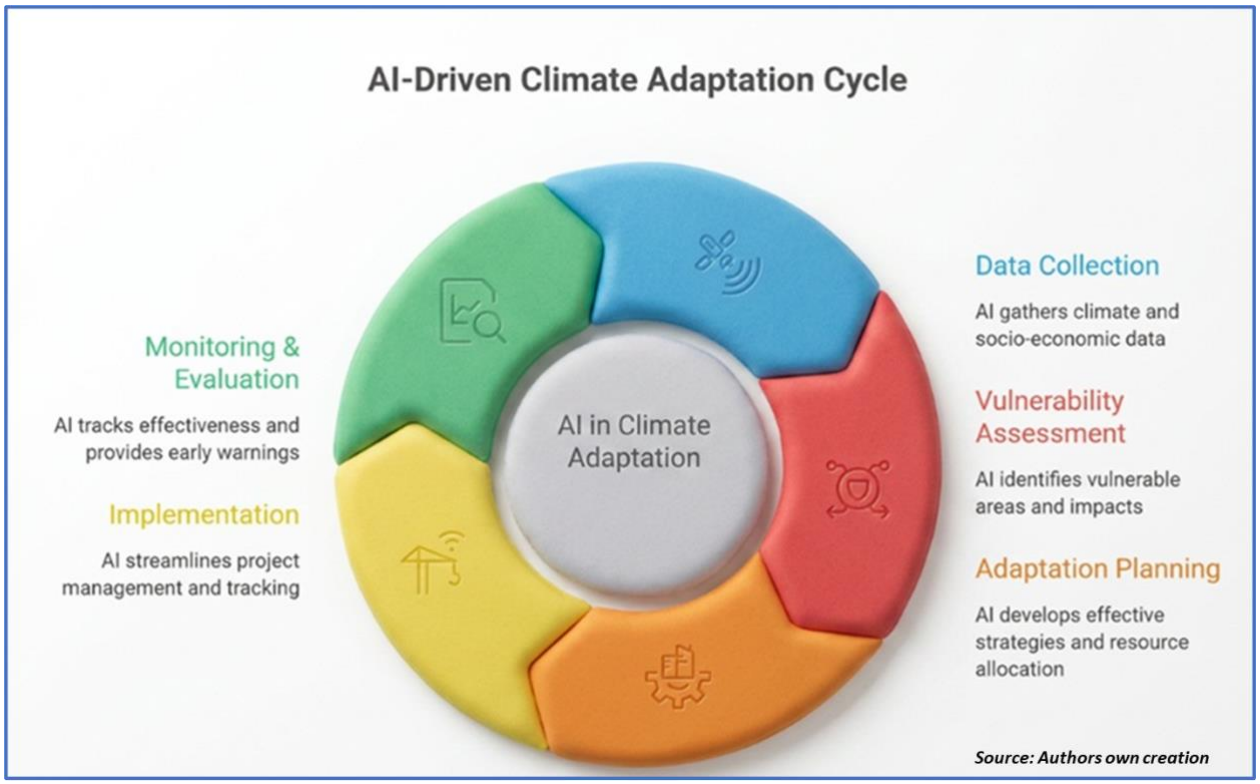
There has been substantial increase in evidence on the use of AI from identifying vulnerabilities, supporting infrastructure planning, ecosystem management, biodiversity conservation, addressing health impacts, sustainable water use, precision agriculture, food

security and land restoration to predicting extreme weather events such as hurricanes, floods, and droughts, enabling proactive disaster risk management, serving as early warning systems, and parameterisation of climate models. AI can be instrumental across the phases of the adaptation cycle - impact, vulnerability, and risk assessment (IVRA), planning and implementation, and monitoring, evaluation, reporting and learning (MERL). This can be achieved through AI assisted data collection, climate modelling, assessment of the pathways, planning, implementation, and monitoring.

AI algorithms can compile, identify the most suitable indicators, and analyse the vast historical climate data, satellite imagery, emissions inventories, socio-economic indicators to determine the current state of the climate system, the drivers and identify areas most vulnerable to climate impacts like flooding, drought, and sea-level rise, perform predictive modelling for the impacts and enhance the ability to forecast medium-range and long-term weather impact on crop production, water resources, infrastructures, and communities' vulnerability under different climate change scenarios.

It can have a significant impact on the implementation and tracking progress of the adaptation project. AI has a huge role in assisting the MEL system for tracking the effectiveness of climate adaptation measures and allowing for adjustments and improvements. Evaluation of the policies in the past and the potential impact of policies before their implementation is another area where AI can be a very useful tool for decision makers.

Figure 1 Use of AI across adaptation cycle



4. Concerns, risk, challenges and trade-offs – AI for adaptation metrics

Growing experience with AI in adaptation has revealed significant concerns, risks, and challenges—both for adaptation generally and specifically for metrics and MERL. The major challenges and risks identified can be categorised into ethical, governance and capacity

I. Ethical challenges - Data quality, security, and privacy risks

a) Availability of high-quality data serves as a major constraint for most of the vulnerable LMICs (Low- and Middle-Income Countries), LDCs (Least Developed Countries) and SIDS (Small Island Developing States) as they may not have the data available to train AI and deploy AI outputs, compared to other regions. This acts as a deterrence and adversely impacts the choice of metrics and the evaluation results from using AI.

b) Use of AI has not been very efficient in climate modelling and long-term scenario simulation for risk-assessment as these are heavily dependent on granularity, coverage, quality of input data and high computational demand. Insufficient or poor data quality specifically on socioeconomic indicators, failure to account for changes in policies and human behaviour can lead to incorrect or misleading outputs. This hinders model accuracy and predictive power, leading to biased assessment of adaptation options.

c) The credibility of the results in assisting decision making may be hindered due to the security concerns associated with use of AI such as vulnerability to cyber-attacks, data breaches, deceptive manipulation of algorithms resulting in

compromised decision and susceptible to issues such as cross-border data sovereignty disputes, national security, and privacy concerns

d) Poor data quality used for training might lead to AI 'hallucination' where the fabricated, distorted, logically inconsistent or incorrect information may be presented as factual output. These false outputs may appear credible and pragmatic. It is more often the outcome of insufficient or biased training data, lack of contextual clarity or flawed data generation methods.

For example, a model might incorrectly identify flood-risk zones and recommend nonexistent adaptation interventions—even proposing metrics to measure these phantom projects. These could have serious consequences such as misdirected resource allocation or decision in favour of potential maladaptation, reduced trust and missed opportunities.

II. Governance challenges – Bias, equity, trust, and accountability issues

a) AI algorithms can perpetuate biases present in the training data, leading to unfair or inaccurate outcomes. For example, adaptation metrics may favour wealthier communities with better data availability and institutional capacity over under-resourced regions leading to weaker adaptation insights and reinforcing existing inequalities in adaptation planning.

b) Regional disparities in adaptation focus and indicator types (e.g., developing countries emphasizing process indicators over outcome metrics) can produce biased AI outputs that undermine decision-making quality.

c) When trained on biased datasets, AI models outputs may overlook women and marginalized communities and fail to address gender and societal inequities.

d) AI might not essentially incorporate the diverse perspectives essential for contextualization and could potentially have adverse impact on vulnerable communities that it aims to make resilient, resulting in biased decisions and potential maladaptation.

e) The risk of AI replacing humans and the subsequent doubts on the credibility of outputs due to inability to understand how the decision was made by the AI system.

f) The output generated by the AI system may be too complex and may have interpretation issues making it challenging to understand how decisions are made; a phenomenon commonly called a "black box,". This lack of transparency can lead to accountability issues.

III. Capacity-building, and Knowledge Sharing

a) Lack of organizational capacity, expertise, knowledge in both AI and climate change sciences, ability for ground truthing or validating the AI outputs, contextualization ability, and capacity to oversee the AI application as an effective tool for augmenting decisions within organizations are persistent gaps in applying AI for adaptation.

b) Communicating the output or knowledge sharing to different stakeholders if not aligned to suit the target audience may serve as a deterrent and hinder the trustworthiness of the AI output.

IV. Trade-offs

a) Transparency and Interpretability: A key trade-off is the conflict between achieving high-performance predictive accuracy with AI models and maintaining transparency and interpretability of those models.
b) Local Context and Standardization: Adaptation metrics cannot be standardized due to the very different local contexts (ecological, sociological, economic) in various places. Incorporating this critical local context rationality into AI systems is difficult.
c) Data Accessibility and Overdependence: While AI can overcome data accessibility hurdles in socio-ecological processes and MERL through supervised data mining, there is a risk of overdependence and over-validation of AI-generated knowledge, especially in low-capacity MERL exercises that require technically rigorous, long-term ground assessments.
d) System Simplicity and Rigor: The temptation of using simple systems that integrate multi-modal data streams is prone to setting up mechanisms that lack transparency and sustained technical rigor.

The key risks, relevance for the non-technical adaptation metrics user and the actions required have been summarized in Table A1.

5. Scope for employing AI by users

This statement defines the proposed boundary for principles governing how deployers and users should employ ready-made AI/ML services in the areas of adaptation metrics and MERL. It deliberately excludes responsibilities belonging to core model creators, foundational-model laboratories, and other deep technical actors. The scope focuses on selecting, configuring, verifying and interpreting tools, not on building or training the underlying models.

Table 2: Scope for the users

Dimension	Scope for users
Primary objective	Selecting, configuring, and operating ready-made AI/ML tools (LLM chat interfaces, SaaS analytics dashboards, API-based prediction services) to generate or interpret adaptation metrics and MEL evidence.
Lifecycle phases	1. Problem definition & indicator selection—identify where an off-the-shelf AI service helps. 2. Vendor assessment & procurement—check documentation, security, privacy posture. 3. Configuration & data ingestion—upload data, set thresholds, define prompts. 4. Verification & validation—sample checks, compare to baselines, monitor drift/bias. 5. Routine operation & interpretation—use outputs in reports and decisions. 6. Periodic review & de-commissioning—decide when to retrain (via vendor) or switch tools.
Stakeholders addressed	Government adaptation planners; multilateral & bilateral donors; NGOs & community-based organisations; climate-adaptation and climate-risk researchers; private developers of adaptation-analytics platforms; evaluation consultants; field practitioners acting as end-users or commissioners of AI-enabled analytics.
Data types covered	Earth-observation imagery; in-situ sensor feeds; socio-economic surveys; project-monitoring records; literature for evidence synthesis; climate-projection model outputs as analysed by the user.
Risk dimensions (user-managed)	Appropriateness of tool; data privacy & consent; misinterpretation, bias, drift; verification of sources & explanations; basic security hygiene (access control, prompt hygiene); accuracy and trustworthiness of AI outputs; energy/compute use at application layer; governance and responsibility for AI outputs.
Decision stakes	Medium- to high-stakes adaptation use cases—fund allocation, public resilience indices, policy decisions.
Time horizon	Immediate to medium term (1–5 years) with biennial review of guidance as vendor offerings evolve.

6. Key Principles for Responsible Use of AI

The integration of AI in climate change adaptation metrics and MERL has a wide array of applications ranging from review of literature and evidence synthesis, remote sensing and geo-spatial analysis, predictive modelling, natural language processing for qualitative data, to real time data analysis for monitoring and evaluation. This requires adherence to following key Principles for responsible use of AI that could serve as a guideline for the adaptation decision makers, practitioners, researchers, project implementers, organizations using AI for metrics and MERL and funders. Table 3 enumerates the Principles along with implementation guidance.

Table 3: AI Principles for responsible use of AI - implementation guidance

Principle cluster	Description	Implementation guidance
Ethical considerations – Data quality, security, and privacy risks		
Human augmentation, not replacement	Ensure that AI is to support rather than replace necessary human expertise and stakeholder input.	Document specific tasks AI will support; identify which decisions remain with humans
Data quality, provenance, accessibility, and traceability	Establish robust data governance frameworks that facilitate data availability, reliability, documentation, and traceability across the AI-supported workflow.	Document data provenance; conduct basic completeness and accuracy checks
Privacy, security, legal compliance, and fallback arrangements	Mechanism for privacy and data security, legal compliance, resilience measures, contingency rules, and fallback plans.	Implement access controls; avoid collecting personally identifiable information when possible; document alternative methods if AI systems become unavailable or unreliable
Governance – Bias, equity, trust, access and accountability		
Human oversight and accountability	Mechanism for human oversight, impact assessment, auditability, and the preference for assisted rather than autonomous intelligence.	Assign named individuals to review all AI outputs before use in decisions
Transparency and explainability	Requires disclosure of AI use, clarity on data sources and decision logic, and explainable AI (XAI) where appropriate.	Disclose to stakeholders that AI was used; identify which outputs are AI-generated; prioritize how AI models make decisions. Train staff to interpret XAI outputs
Participation, local knowledge, and ground-truthing	Involve stakeholder engagement, indigenous and local knowledge, validation, feasibility, and social acceptability.	Conduct stakeholder validation sessions before finalizing AI-based metrics or assessments
Fit-for-purpose tool selection and configuration	Selection of AI tools and settings that match the specific adaptation task, local context, scale, and capacity.	Choose from the available or customize (if required) AI tools that consider local context, data availability, and capacity requirements.
Bias, inequity, and exclusion risks	Ensure identification and mitigation of bias, inequity, and demographic or regional exclusion.	Test AI outputs across different demographic groups; report performance gaps publicly
Validation and ongoing monitoring of AI-supported outputs	Evaluation of AI-supported metrics, performance monitoring, and review of unintended outcomes over time.	Choose tools with explanation features; Train staff to interpret XAI outputs
Capacity - Building Capacity and Knowledge Sharing		
Capacity, training, and internal guidance	**Advance AI literacy, staff capability, and internal operational guidance for responsible use.**	**Facilitate training and skill development programs for M&E professionals, climate scientists, and other stakeholders including local communities to advance AI literacy and ensure they can effectively utilize and interpret AI-generated insights.**
Clear communication for different audiences	Explanation of AI-supported findings in forms suited to policymakers, practitioners, researchers, and communities.	Develop audience-specific communication plans. For example, technical explanations for data scientists or researchers and a simpler explanation for policymakers and community members
Practical implementation guidance and shared learning	Involves detailed guidance, collaboration, knowledge exchange, and dissemination of lessons learned.	Establish collaborative frameworks involving international partnerships and cooperative frameworks to facilitate knowledge exchange, technology transfer, and capacity-building, in line with the provisions of the UNFCCC and the Paris Agreement.

Note: The shaded rows refer to the '**Essential' Principles** that are non-negotiable minimum standards, hence mandatory. The remaining are **'Important' Principles** that strengthen quality and trustworthiness and represents the best practices for well-resourced set ups.

7. Conclusion and Way forward

The Principles enumerated in this document are essentially general, broad, and overarching standards that serve as a foundational guidance on responsible use of AI for adaptation metrics and MERL. These Principles have been confirmed following the inputs from the respondents of the online survey. Based on the feedback received, among the 'essential' Principles, 'Data quality, provenance, accessibility, and traceability' would be the hardest to implement followed by establishing mechanisms for 'human oversight and accountability.' The prominent areas that deserve more attention include credibility and trustworthiness of AI-supported outputs, data quality, provenance, and representativeness, human oversight and accountability and Local contextualization and ground-truthing.

As a way forward we suggest that while implementing the AI principles, the deployers should be cautious and make efforts to establish robust data governance frameworks that facilitate data availability, reliability, documentation, and traceability across the AI-supported workflow. Additional efforts should be made on having a robust mechanism to ensure human oversight, impact assessment, auditability, and the preference for assisted rather than autonomous intelligence. Further, the dynamic nature of the AI system needs to be profoundly contemplated as with time these principles for responsible use of AI may lose relevance. Keeping in with the dynamic nature of AI, the principles need to be regularly reviewed and refined in consultation with the stakeholders.

Deliberation Next Phase: options for the practical application of the Principles

The Principles provide a common foundation for responsible use of AI in adaptation metrics, monitoring, evaluation, reporting and learning. However, they will only influence practice if users understand how to apply them in real settings. IPAM should therefore support further deliberation on how the Principles can be made practical, credible, and proportionate for different users and contexts.

This next phase should not presume that IPAM will develop a formal compliance, certification, or assurance mechanism. Instead, IPAM should examine how far it can and should go in supporting responsible application, considering its mandate, capacity, governance role, resources, and the needs of the adaptation metrics community. The aim should be to identify what forms of guidance, documentation, due diligence, testing, or tools may be useful before deciding whether any formal mechanism is appropriate.

This work is important because AI-supported outputs may increasingly inform adaptation indicators, evaluations, monitoring reports, vulnerability assessments, policy appraisal, public reporting, and investment decisions. Users need to know not only that AI was used, but how it was used, what checks were applied, what limitations remain, and how human judgement and contextual validation were maintained. Without further guidance, the Principles could be cited too loosely or treated as too general to guide responsible practice.

IPAM shall therefore consider the following areas for further deliberation:

1. Consider whether a minimum due diligence threshold is needed for use of the Principles.

IPAM should deliberate on what it would mean for a practitioner, organization, funder, evaluator, or policymaker to claim that AI has been used in accordance with the Principles. Options could range from simple self-declaration, to disclosure supported by a basic evidence record, to more structured forms of peer review, assurance, or certification.

The purpose would be to determine what level of due diligence is useful, proportionate, feasible, and credible. A light approach may be accessible but provide limited assurance. A formal approach may provide greater confidence but would raise questions about IPAM's role, capacity, liability, and institutional readiness. IPAM should therefore avoid assuming a certification or compliance role before these implications are examined.

A practical starting point may be to consider whether any claim of alignment should be accompanied by a minimum evidence record documenting the purpose of AI use, tools used, data sources, human oversight, validation checks, bias and contextual considerations, known limitations, and disclosure of AI use.

2. Consider whether practical guidance is needed for applying the Principles in different contexts.

IPAM should assess where users would benefit from more specific guidance. AI risks are not uniform. The safeguards needed for drafting, summarisation, or exploratory literature review differ from those needed for indicator selection, project evaluation, vulnerability scoring, policy appraisal, adaptation finance assessment, or public reporting.

This work stream would examine whether context-specific guidance should be developed for priority applications, such as project evaluations, indicator selection, evidence synthesis, review of project or policy reports, adaptation finance or investment

assessment, and communication of adaptation progress. For each context, guidance could identify what should be avoided, what minimum checks should be applied, and what good practice looks like.

Such guidance would help users distinguish acceptable AI-assisted professional judgement from uses that are insufficiently transparent, inadequately validated, or likely to create risks of bias, misinterpretation, or maladaptation.

3. Consider whether prompts, checklists, or decision-tree tools could support practical due diligence.

IPAM should explore whether structured tools could help practitioners apply the Principles consistently before, during, and after AI use. Prompts, checklists, or decision-tree tools could help users ask the right questions about task suitability, data inputs, verification needs, human review, bias and contextual risks, and communication of limitations.

These tools could also help users prepare a practical record of AI use, making it easier to demonstrate that reasonable due diligence was applied. They should be framed as aids to professional judgement, not substitutes for expertise, stakeholder engagement, ground-truthing, or independent validation.

4. Define use cases that can serve as test beds for applying the Principles.

IPAM should identify a small number of practical use cases through which the Principles, due diligence options, guidance materials, prompts, checklists, and decision-tree tools can be tested. These use cases should represent different levels of risk, technical complexity, data sensitivity, and decision significance. The purpose would not be to endorse particular AI tools or outputs, but to understand how the Principles operate in realistic adaptation metrics and MERL workflows.

Suitable test-bed use cases could include AI-assisted review of project or policy reports, indicator selection for adaptation monitoring frameworks, evidence synthesis for evaluations, assessment of project contributions to resilience outcomes, or review of adaptation finance and investment claims. More technical use cases, such as remote sensing, vulnerability scoring, predictive modelling, or automated monitoring, could be considered once simpler cases have been tested.

Together, these four work streams would allow IPAM to move carefully from broad principles toward practical application. They would help determine how far IPAM can and should go in supporting responsible AI use, while preserving the distinction between guidance, assurance, and certification. The next phase should therefore focus on consultation, learning, testing, and deliberation before IPAM decides whether to develop any formal mechanism.

Appendix A1

Table A1: Key AI Risks/Impacts and Scope/Relevance for Adaptation Metrics Users

Term	Concise definition	Why it matters for adaptation metrics	Practical relevance & advice for non-technical experts	Key action stage(s)	Signals & hand-off triggers for non-technical staff
Jailbreak defence	Layered controls that keep an LLM from being tricked into breaking its own rules.	Prevents leakage of sensitive geo-data and fabricated scores.	Ask vendors for refusal-rate metrics and red-team results.	Pre-procurement; operational monitoring	*Red flag:* model suddenly answers forbidden questions or produces extremist content. → **Pause use and call AI/security lead.**
Prompt injection	Inputs that smuggle hidden instructions to hijack an LLM.	Malicious text in feeds could bias dashboards.	Moderate or strip untrusted text; keep human review for critical outputs.	Daily use	*Symptom:* model output references strange system-level commands or disowns prior instructions. → **Escalate to technical team.**
Retrieval-Augmented Generation (RAG) & vector poisoning	Corrupting the document store a model retrieves from.	Can skew evidence toward denialist or out-dated studies.	Spot-check retrieved docs for source credibility.	Pre-procurement; periodic audit	*Clue:* surge in citations to unknown or suspicious sources. → **Request data-quality audit.**
Red teaming	External stress-testing to find model weaknesses.	Exposes hidden errors before public release.	Require summary of latest red-team report.	Pre-procurement; annual retest	*Missing doc:* vendor can't produce a recent red-team summary. → **Flag for procurement hold.**
Reinforcement Learning from Human Feedback (RLHF)	Fine-tuning guided by human preference labels.	Narrow label set may embed regional bias.	Check who supplied feedback; offer local examples.	Model development / procurement	*Gap:* vendor cannot explain label demographics. → **Ask for diversity review or add local experts.**
Model card	Standard disclosure of purpose, data, metrics, limits.	Enables auditors to judge fitness for context.	Use as checklist (coverage, bias, update cadence).	Pre-procurement; governance docs	*Alert:* no model card or missing sections. → **Delay deployment until provided.**
Differential privacy	Adds noise so outputs don't reveal individuals.	Allows safe use of household surveys.	Confirm DP before sharing granular data.	Data-sharing design	*Warning:* legal/compliance team flags privacy concerns or survey participants express worry. → **Engage privacy specialist.**
Data drift	Real-world data move away from what the model was trained on.	Rainfall shifts can break flood-risk models.	Agree accuracy thresholds; set drift alerts.	Operational monitoring	*Indicator:* dashboard accuracy suddenly drops or alerts fire. → **Schedule retraining or consult data scientist.**
Adversarial robustness	Resistance to deliberately tampered inputs (e.g., tweaked satellite pixels).	Attacks could hide illegal land-use change.	Ask imagery vendors for robustness tests; cross-validate with second data source.	Pre-procurement; usage spot-checks	*Hint:* imagery analysis disagrees sharply with field inspections. → **Commission forensic check.**
Explainable AI (XAI)	Methods that make model reasoning intelligible.	Helps justify why an index spikes in one district.	Choose tools that show plain-language explanations.	Pre-procurement; usage reviews	*Signal:* model gives decision but "no explanation available." → **Escalate and request XAI output.**
Bias mitigation	Techniques to detect and reduce systematic error.	Prevents finance from being misallocated.	Demand demographic performance reports; run fairness review.	Pre-procurement; periodic monitoring	*Sign:* persistent performance gap for a region or group. → **Trigger fairness audit.**
Systemic risk (EU AI Act)	Extra duties for very large general-purpose models.	Non-compliance can stall deployment.	Confirm risk status; obtain mandatory assessment.	Pre-procurement; compliance check	*Cue:* regulatory team asks for Article 51 paperwork that vendor hasn't supplied. → **Escalate to legal/AI compliance.**
Energy footprint	Electricity / carbon cost of running AI.	High compute demand may clash with climate goals.	Track GPU-hour or kWh budgets; prefer smaller fine-tunes.	Pre-procurement; operational reporting	*Observation:* compute bills or energy-use reports spike unexpectedly. → **Ask engineers to optimise model or schedule.**