Available Online: 24 September 2026 DOI: 10.54254/2977-5701/2026.37392

Digital tools or digital facades? institutional conditions for credible AI-enabled carbon regulation in China's multi-level governance

Qinyu Yu

Harris School of Public Policy, The University of Chicago, Chicago, USA

13805861801@163.com

Abstract. As China digitizes carbon governance, the central challenge is shifting from information scarcity to the credibility of digitally generated evidence. This paper examines the institutional conditions under which AI-assisted screening and auditing can strengthen verification and enforcement in a multilevel governance system, as well as how strategic compliance may weaken these effects. It conducts a structured qualitative comparison of Zhejiang, Hebei, and Inner Mongolia, combining cross-case analysis with process tracing based on regulations, government reports, platform documentation, enforcement materials, and investigated cases. Regulatory credibility is evaluated along three dimensions: information integrity, measurement validity, and enforcement closure. The findings show that digital systems have improved data integration, reporting frequency, record traceability, information retrieval, and anomaly screening. However, publicly available evidence is considerably stronger for information acquisition and screening than for the final resolution of individual cases. One official investigation provides direct evidence of reporting-time manipulation and record fabrication, while indicator substitution remains an institutional vulnerability rather than a documented practice. The available evidence does not substantiate platform capture. The study therefore neither estimates the independent causal effect of AI nor ranks the three provinces by regulatory credibility. Instead, it argues that AI-enabled regulation becomes credible only when technical outputs are connected to evidence beyond the control of reporting entities, traceable revision histories, valid indicators, independent verification, and documented final dispositions. The paper concludes by recommending auditable data-provenance standards, institutional separation between algorithmic detection and verification, and standardized public reporting on enforcement closure.

Keywords: AI-enabled regulation, carbon emissions governance, multilevel governance, regulatory credibility, strategic compliance

1. Introduction

Credible carbon regulation depends on regulators' ability to verify how emissions data are produced and whether identified problems are ultimately corrected. In a multilevel system, central authorities depend on data generated by enterprises and reviewed by local and provincial agencies. China's national framework requires

enterprises to submit emissions data and supporting materials, retain electronic verification records, and refer anomalies to local authorities for investigation [1, 2]. Official documentation also describes rule-based screening, parameter relationships, knowledge graphs, and AI-assisted information retrieval. Recent research using data from China's national ETS shows that machine-learning screening can identify potential misreporting and help target limited verification resources [3]. These systems broaden the information available for review, but technology alone cannot establish that source data are authentic, that indicators validly represent policy objectives, or that identified problems have been fully resolved.

Existing research identifies both regulatory improvements and strategic adaptation. Automated monitoring and direct data transmission can reduce local control over regulatory information. After Chinese cities adopted automated air-pollution monitoring, reported PM10 concentrations rose by approximately 35 percent, while satellite aerosol measures showed no comparable increase [4]. Yet measurements that carry material consequences also create a risk of reactivity: regulated actors may reorganize their behavior around what the system observes [5]. Pollution improvements concentrated near monitoring stations and before evaluation deadlines suggest that genuine local gains may not translate into broader, sustained change [6]. Fragmented authority and target-based accountability make this problem especially salient in China [7-9]. Research on public-sector AI likewise tends to document system deployment more often than effects on decisions or outcomes [10]. The central question, therefore, is which institutional conditions allow improved information to generate credible regulatory action. Taken together, these studies imply that monitoring technology changes the opportunity structure of compliance rather than simply removing discretion. Some forms of manipulation become harder, while incentives may shift toward less visible margins of the measurement system.

This paper asks two questions. Under what institutional conditions can AI-assisted identification and auditing in China's multilevel carbon-regulation system produce credible verification and enforcement? Under what conditions can strategic compliance weaken those effects? Regulatory credibility requires verifiable judgments, traceable responsibility for corrective action, and documented final dispositions. AI deployment is therefore the object of analysis, not evidence of regulatory success. The paper argues that AI-enabled tools can reduce information asymmetry, but their contribution to credible regulation depends on four institutional linkages: access to evidence beyond the control of reporting entities; traceable revision histories; valid connections between indicators and policy objectives; and documented enforcement closure. These linkages connect the paper's three evaluative dimensions to observable institutional arrangements. Independent evidence supports information integrity, valid indicator-policy relationships support measurement validity, and documented dispositions establish enforcement closure.

The paper compares Zhejiang, Hebei, and Inner Mongolia within a common national reporting and verification framework. It does not treat them as low-, medium-, and high-credibility cases. Instead, each provides a different analytical entry point. Zhejiang has institutionalized a consequential carbon-efficiency indicator, whose design and effects have been examined empirically [11]. Hebei-based research documents the integration of digital carbon management with steel-sector production-data flows [12], while official provincial materials describe a wider recurring reporting and verification structure. Research on coal-power carbon measurement explains why traceability of coal consumption and key parameters matters [13], while regional reporting documents Inner Mongolia's newer physical-testing capacity. A publicly disclosed fraud incident, also discussed in scholarship on transparency in China's national ETS, provides direct evidence of data-quality risks [14]. The analysis draws on official regulations, government reports, platform guidance, investigation records, and enforcement materials, using within-case process tracing alongside structured cross-case comparison. Observed conduct is assessed separately from institutional design, and undisclosed information is coded as unknown rather than as evidence of regulatory failure.

The findings show meaningful but uneven institutional change. Digital platforms have improved data integration, reporting regularity, searchability, and anomaly screening. One investigated incident provides direct evidence of reporting-time manipulation and record fabrication [14]. Because consequential composite indicators can narrow regulatory attention, indicator substitution is an institutional vulnerability; the available evidence does not, however, show that actors manipulated Zhejiang's carbon-efficiency indicators. Nor does it substantiate platform capture. At the national level, published review volumes, declining issue rates, and follow-up spot checks show that the review system is operational [3], but public documents do not connect individual algorithmic flags to case-level outcomes. The paper therefore neither estimates the independent causal effect of AI nor ranks the provinces. Its central finding is conditional: technological visibility strengthens regulatory credibility only when an auditable chain connects source evidence to final disposition.

The paper contributes to research on algorithmic regulation by treating AI as one component of a sociotechnical regulatory system rather than as an autonomous technological intervention [15, 16]. It also refines principal-agent and reactivity accounts: digitization may constrain discretion at one stage while relocating it to data selection, indicator design, verification, or follow-up.

2. Theoretical logic

2.1. Digital monitoring as an information solution

Digital monitoring can address a classic principal-agent problem. Central authorities set targets, but enterprises and local officials possess superior information about production, energy use, emissions, and implementation. Compliance assessments therefore depend partly on information generated or transmitted by the regulated agents themselves. In China, fragmented authority, hierarchical personnel incentives, and target-based accountability further intensify this asymmetry by allowing local actors to influence calculation and reporting practices [7-9]. Apparently precise indicators may thus conceal substantial discretion in the processes that generate them. The principal's problem is therefore not only insufficient information, but also dependence on agents for the categories, records, and calculations through which compliance becomes legible.

AI and other digital tools can reduce this asymmetry by accelerating anomaly detection, standardizing procedures, and widening access to information. Public-sector AI, however, is often evaluated by whether a system has been deployed rather than by whether it changes decisions or outcomes [10]. The relevant question is therefore whether technology changes who controls regulatory information and who can verify it.

Greenstone et al. identify one such mechanism. After Chinese cities adopted automated monitoring, reported PM10 concentrations rose immediately by approximately 35 percent, whereas satellite aerosol data showed no corresponding inflection [4]. Direct transmission and centralized quality control constrained local discretion and enabled verification by higher-level authorities. This finding does not establish that digitization is inherently trustworthy. Its regulatory value still depends on control over raw data, traceable revisions, and independent verification.

Proposition 1: Digital monitoring is more likely to enhance the credibility of information when it restricts local discretion over data transmission, preserves higher-level authorities' ability to access raw records, and allows for verification independent of the reporter.

2.2. Reactivity and strategic adaptation

Closing one channel of manipulation does not eliminate strategic behavior. Public measurement is reactive: once an indicator carries tangible consequences, those being measured adjust their behavior to it [5]. Indicators can become operational targets in their own right, encouraging actors to improve visible performance without

producing commensurate system-wide gains. Adaptation may take the form of data falsification or of concentrating genuine effort within the spatial, temporal, or organizational boundaries visible to the monitoring system.

Yang et al. illustrate this dynamic. Pollution fell approximately 3.2 percent more within three kilometers of automated monitoring stations than in more distant areas, and the difference widened as the evaluation deadline approached [6]. Direct data transmission constrains conventional manipulation of readings, but a fixed monitoring network can also encourage authorities to concentrate improvements around stations rather than across the city. Such local improvements may be real while remaining misaligned with system-wide objectives.

China's target-responsibility system produces comparable temporal and reporting responses. Local governments have met energy-efficiency targets through temporary production suspensions and emergency power cuts [17]. Corporate records also reflect administrative error, intentional underreporting, delayed submission, and calculation choices influenced by local leaders [9]. Capacity failure should be distinguished from reporting manipulation, which alters records, and from selective substantive compliance, which changes real behavior only where or when evaluators can observe it. Fiscal and political incentives may also affect disclosure and decision-making, although the available evidence establishes correlation rather than causation [18].

This paper defines performative compliance as the production of data or visible performance that satisfies formal oversight requirements without achieving policy objectives at a commensurate scale, for a sustained period, and across the relevant system. It may occur through gaming of metrics or measurement boundaries, temporal or reporting manipulation, or control over the digital channels through which performance becomes visible to regulators. The definition does not treat every response to an indicator as problematic. Adaptation becomes performative when formal performance is decoupled from the scope, duration, or substance of the underlying policy objective.

Proposition 2: When evaluations rely on narrow, predictable, and high-stakes metrics, enhanced monitoring may shift strategic behavior to the boundaries of metrics, space, or time, rather than eliminating it.

2.3. From better information to credible regulation

Accurate information alone does not produce credible regulation. Yeung conceptualizes algorithmic regulation as a feedback process linking standard setting, information collection, decision-making, intervention, and evaluation [15]. A break at any point can prevent technological insight from generating regulatory change: accurate data may be translated into invalid proxies, anomalies may lack an accountable reviewer, or formal enforcement may proceed without verification that corrective action occurred.

Regulatory credibility is therefore evaluated along three connected dimensions. Information integrity concerns whether raw observations are authentic, preserved, and traceable through later revisions. Measurement validity concerns whether metrics and model outputs correspond to the substantive behavior that policy seeks to change. Enforcement closure concerns whether identified problems receive institutional review, lead to enforceable action, and culminate in verified correction. AI can support the first two dimensions through anomaly detection and cross-validation, but it cannot supply the authority, organizational capacity, or incentives required for the third. These dimensions are analytically distinct. Authentic source records can still feed a poorly designed indicator, and a valid alert can still fail if no institution is responsible for investigating and closing the case.

Digital platforms are themselves sociotechnical institutions. Data definitions, thresholds, permissions, model updates, procurement arrangements, and human review jointly determine what a system can reveal [15,

16]. A claim of platform capture therefore requires evidence that actors with conflicts of interest can alter, filter, suppress, or restrict access to relevant information. Outsourcing or local operation alone does not establish capture; evidence of actual authority, decisions, or information flows is necessary.

Proposition 3: Only when anomalies can be independently reviewed, assigned to accountable decision-makers, linked to actionable remedies, and subject to verifiable follow-up on corrective actions can better information be transformed into trustworthy regulation.

Taken together, digitization can close existing reporting loopholes while relocating discretion to indicator design, monitoring boundaries, or enforcement. The analysis therefore examines institutional safeguards and vulnerabilities in the three provinces without imposing a performance ranking or treating nondisclosure as proof of failure.

3. Methodology

3.1. Research design and case selection

This study uses an embedded, document-based comparative design. The national system for regulating carbon-emissions data constitutes the overarching case, and Zhejiang, Hebei, and Inner Mongolia are provincial implementation settings within it. A structured, focused comparison applies the same theoretical questions to each setting [19]. The unit of analysis is not the province as a whole, but the provincial implementation of carbon-data reporting, digital screening, technical verification, and regulatory follow-up.

The provinces were selected purposively to capture variation in industrial structure, energy dependence, administrative conditions, and publicly observable arrangements for digital regulation. This diverse-case strategy identifies theoretically relevant contrasts rather than a representative sample [20]. Selection was not based on a prior ranking of regulatory credibility or presumed low, medium, and high reporting gaps, because comparable independent evidence is unavailable. The study draws bounded inferences about how a shared national framework operates under different institutional conditions; it does not estimate the average effect of AI across provinces. The cases are therefore used for analytical leverage rather than statistical generalization. Similarities can reveal common constraints within the national architecture, while differences help identify which institutional linkages vary across implementation settings.

3.2. Data sources and evidence evaluation

The study relies primarily on publicly available documentary evidence: regulations, technical specifications, procurement and acceptance records, platform descriptions, verification and inspection records, penalty decisions, rectification notices, and official reports. Academic research supplies theoretical, contextual, and methodological foundations. Supplier and media materials are used to identify technologies or events, but central claims are corroborated with independent sources wherever possible. The documentary corpus is especially useful for reconstructing formal authority, mandated workflows, and publicly acknowledged violations. It is less capable of revealing routine informal practice or internal disagreements that leave no public record.

Evidence is assessed in relation to the claim it is used to support. Audits, penalty decisions, original records, and a complete discovery-review-disposition chain offer the strongest evidence of actual conduct. Regulations, technical specifications, procurement records, and process descriptions establish formal capacity or institutional vulnerability, but not actual operation. Aggregate statistics and proxies may reveal anomalies without establishing strategic intent, and promotional materials cannot demonstrate implementation or effectiveness. Reposts of the same announcement are treated as a single source. This claim-specific approach

avoids assigning a fixed credibility score to an entire source. An official regulation may be authoritative evidence of a legal requirement but weak evidence that the requirement was implemented in a particular case.

Undisclosed enforcement outcomes are coded as unknown, and the existence of a platform or review process is not treated as evidence of greater credibility. The study also distinguishes AI from digitalization more broadly. A system is coded as AI-enabled only when its documentation expressly identifies algorithmic functions such as machine-learning prediction, computer vision, natural-language processing, or model-based anomaly detection. Platforms, sensors, and fixed-threshold alerts are classified as digital or automated monitoring [10, 15]. The coding rule is deliberately conservative because official materials often use terms such as intelligent, smart, or digital without identifying the underlying technique. Treating those labels as proof of AI would blur the paper's main analytical distinction.

3.3. Analysis procedure

Each provincial setting was examined through five common questions. What regulatory tasks does the technology perform? What data do it receive and produce? Who can access, modify, or aggregate those data? How are anomalies reviewed? Do public records document subsequent decisions, corrective action, or verification? The answers were then evaluated against the three dimensions developed above: information integrity, measurement validity, and enforcement closure. Applying the same questions to all three settings creates a common evidentiary template even where the available records differ in depth. It also makes missing links visible without treating every missing disclosure as evidence of nonperformance.

The three forms of performative compliance are treated as candidate mechanisms, not predetermined findings. Evidence is classified as supported, consistent but not demonstrative, institutional vulnerability only, or insufficient. Process tracing is used only where the documentary record permits reconstruction of a specific event, and its diagnostic value is assessed against candidate mechanisms and alternative explanations [21]. A penalty imposed after a platform's launch is attributed to that platform only when the record links a system signal to the investigation. Equipment failure, capacity constraints, legitimate correction, industrial adjustment, and general enforcement activity are considered alternative explanations. The categories distinguish positive evidence of conduct from evidence that a system creates incentives or opportunities for that conduct. This distinction is essential when public records disclose architecture more often than behavior.

3.4. Limitations

This design cannot observe every warning, internal model output, data revision, or enforcement decision. Public disclosure is selective and varies across provinces, producing uneven evidentiary depth. The study therefore cannot calculate alert-confirmation rates, rank the three provinces by regulatory credibility, or estimate AI's causal effect on emissions or enforcement. Its broader conclusions concern institutional design and publicly observable auditability; stronger claims about mechanisms are confined to events supported by direct evidence. Because no interviews were conducted, interviews are not part of the evidence base. Selection effects are also possible because authorities are more likely to publicize successful initiatives or serious violations than routine cases with ambiguous outcomes. The study consequently treats the public record as evidence of what can be observed, not as a complete census of regulatory practice.

4. Analysis

4.1. Digital regulatory framework and data integrity

China's National Emissions Trading System (ETS) combines centrally established rules with decentralized implementation. The *2024 Interim Regulations on the Administration of Carbon Emissions Trading* assign overall supervisory responsibility to the Ministry of Ecology and Environment and require provincial ecological and environmental authorities to compile annual lists of key emitters, allocate allowances, and verify annual emissions reports. Key emitters remain legally responsible for the authenticity, completeness, and accuracy of their data and reports. Provincial authorities may commission technical service agencies to conduct reviews, but those agencies must operate independently and bear separate liability for false or materially defective results. The Regulations also authorize on-site inspection and prescribe distinct penalties for enterprises and technical service agencies that falsify data [1]. Formal centralization has therefore not created a single central monitoring authority. Measurement, verification, and enforcement remain distributed among enterprises, contracted agencies, and several administrative levels.

For the power-generation sector, a 2023 notice from the Ministry of Ecology and Environment specifies the data-management process in greater detail. Provincial authorities organize the annual inventory, data-quality control plan, monthly digital-evidence archive, annual reporting, and verification. Within forty calendar days after each month ends, enterprises must upload data on fuel consumption, lower heating value, elemental carbon content, electricity and heat generation, operating hours, load factor, and supporting materials. Annual reports then undergo documentary review, on-site verification, and online submission, with electronic records retained as required. Technical reviewers screen monthly submissions for anomalies. Identified problems are referred to municipal ecological and environmental authorities for verification and rectification, and the resulting recommendations return to provincial authorities through the national platform [2]. This Monitoring, Reporting, and Verification (MRV) framework improves traceability but is not equivalent to continuous, independent, direct emissions measurement. Much of the platform's input still originates in enterprise records and sampled fuel properties.

The national platform adds a unified digital layer to this decentralized governance chain. The Ministry of Ecology and Environment describes a closed loop of pre-submission standardization, in-process intelligent warning, and post-submission checklist verification. Its validation model applies more than 100 rules concerning theoretical limits, thresholds, standard values, and logical relationships, while a knowledge graph connects reporting, verification, trading, and compliance data to support anomaly detection. The same report states that AI assists regulators in retrieving carbon-market information; it describes anomaly screening more broadly as a combination of rules, models, parameter relationships, and knowledge-graph links [3]. This paper therefore does not classify every automated alert as AI. Official materials report that in 2024 the platform reviewed more than three million power-sector data entries through a three-tier national-provincial-municipal mechanism. Issues identified in provincial and municipal reviews fell by approximately 30 percent, noncompliant items in verification reports fell by about 24 percent, and five unannounced inspections targeted anomalous composite indicators or inadequate rectification [3]. These figures establish operational activity, but aggregate national trends neither identify which tools caused the changes nor show that every alert reached enforcement closure. The decline in identified problems is compatible with improved compliance, better guidance, changing review practice, or changes in the composition of reported data. Without case-linked records, the aggregate figures cannot distinguish among these explanations.

Provincial materials show how diverse local systems overlay this national template. In Zhejiang, a National Data Administration presentation on Huzhou's Carbon Efficiency+ system describes a data warehouse

integrating records from multiple departments, electricity-to-carbon conversion factors, dynamic warnings, and high-precision sensors at large coal-fired power plants. The system reportedly aggregates 11 million energy-consumption records to support monitoring across 41 industries and 11 prefecture-level jurisdictions, while the industrial carbon-efficiency code covers 49,345 above-designated-size enterprises [11]. This coverage improves temporal resolution and cross-source comparison, although emissions for most enterprises must still be estimated from activity data and conversion factors. Public materials do not disclose model error rates, revision permissions, or independent validation results. The Zhejiang evidence therefore establishes extensive data integration and institutionalized use of indicators, but not independent verification or enforcement closure. Zhejiang therefore illustrates both the administrative value and the validity risk of integrated indicators. Combining dispersed records can reveal patterns unavailable to a single agency, yet each conversion factor and weighting choice introduces assumptions that require validation.

Hebei places greater emphasis on reporting frequency and administrative coverage. A 2025 response by the Hebei Provincial Department of Ecology and Environment to a provincial People's Congress proposal states that the comprehensive carbon-emissions supervision platform covers more than 400 enterprises in power generation, steel, cement, glass, chemicals, and petrochemicals, with daily reporting, monthly archiving, and annual verification [12]. This structure creates recurring records that can support longitudinal consistency checks. The response does not specify which variables are reported daily, whether instruments transmit data directly, how historical revisions are logged, or how many alerts have been confirmed. The evidence thus establishes an operating institutional framework, but not independent data generation or improved enforcement attributable to its alerts. Frequency has regulatory value when repeated observations allow reviewers to locate discontinuities, late revisions, or implausible relationships across time. Its value is weaker when repeated reports originate from the same unverified calculation or are freely amendable without a version history.

Inner Mongolia is distinctive for its developing physical-measurement capacity. Regional reports state that the National Carbon Measurement Center (Inner Mongolia) completed a laboratory by the end of 2024. The laboratory can test more than 50 parameters, including proximate coal-quality measures and elemental indicators, and has inspected 74 batches of coal samples from key emitting entities [13]. Such testing can independently verify fuel parameters reported by enterprises, a recognized vulnerability in power-sector emissions calculations. The available materials, however, do not describe the sampling frame, the number of discrepancies or corrected reports, penalty outcomes, or how laboratory findings enter the national platform. They demonstrate verification capacity rather than verified enforcement outcomes. Physical testing is particularly important because it introduces an evidentiary source that is not simply another transformation of enterprise-reported activity data. Its independence nevertheless depends on sampling, custody, laboratory governance, and linkage to subsequent decisions.

This evidence also changes the basis for interprovincial comparison. Visible platform coverage cannot be read directly as regulatory credibility. A broad dashboard may still depend on modeled proxies; frequent reporting may reproduce errors from a common source; and a capable laboratory may test only selected samples. The relevant comparison is therefore not technological sophistication in the abstract, but the extent to which each system permits verification independent of the actors that supply compliance data.

The three provinces have adopted complementary but incomplete approaches to information integrity. Zhejiang emphasizes multi-source integration and proxy-based assessment, Hebei increases reporting frequency, and Inner Mongolia strengthens physical testing of selected parameters. None of the public records provides a complete chain from raw inputs and revision history through anomaly detection, verification, correction, and final disposition. Digital systems may have narrowed the scope for untraceable alteration, but

they have not removed institutional dependence on provincial capacity, the independence of contracted agencies, municipal investigation, and verifiable corrective action. These conditions determine whether digital signals become credible regulatory evidence.

4.2. Uneven evidence for strategic adaptation and performative compliance

The evidence does not support the three candidate mechanisms equally. Applying the standards set out in Section 3 produces a deliberately asymmetric result: one investigated incident directly demonstrates reporting-time manipulation; metric and measurement-boundary gaming remain institutional vulnerabilities; and platform capture is not substantiated. This result is more defensible than assigning one mechanism to each province because it separates observed conduct from behavior that an institutional arrangement merely makes possible. The asymmetric finding also protects the analysis from an ecological fallacy. Evidence concerning one enterprise, contractor, or investigated event cannot automatically characterize a province, just as a platform's formal capability cannot establish the conduct of its users.

Zhejiang offers the clearest institutional conditions for metric reactivity, but not evidence that such reactivity has occurred. Huzhou's Carbon Efficiency+ system converts electricity consumption and other administrative data into emissions estimates, then combines emissions levels, carbon-use efficiency, and progress toward carbon neutrality in a composite carbon-efficiency score. The metric has practical consequences: public cases state that the code informs Green Factory star ratings, performance-per-mu evaluation, and credit approval [11]. These consequences create an incentive for evaluated enterprises to improve their scores and therefore satisfy a necessary condition for reactivity. They may also create scope for measurement-boundary gaming if enterprises can relocate production, shift the timing of recorded activity, or optimize only the variables included in the index. Public records, however, provide neither firm-level before-and-after data nor indicator weights, audit findings, or documented cases of such adaptation. The evidence therefore establishes an institutional vulnerability, not indicator substitution. Inferring substitution solely from the indicator's practical consequences would exceed what the record supports. More direct evidence would require observable changes clustered around score thresholds, internal documents describing strategic responses, or audits showing divergence between score improvement and broader emissions performance. None is available in the reviewed materials.

The strongest direct evidence concerns reporting-time manipulation. In 2022, the Ministry of Ecology and Environment published findings from special supervision conducted by 31 working groups between October and December 2021. The inquiry examined coal-sample collection and preparation, coal-quality testing, data verification, and report preparation by technical service agencies and regulated power-sector enterprises [14]. It found that Zhongtan Energy Investment Technology used editable templates to alter submission dates, test dates, report dates, and report numbers in test reports for regulated enterprises, including Inner Mongolia Ordos High-Tech Materials Co., Ltd. The firm recast bulk sample submissions as monthly submissions and tests, removed QR codes from original reports, falsified elemental-carbon results in some cases, and fabricated total-moisture data used in conversion calculations [14]. These were documented alterations to evidentiary records, not anomalies inferred from an unusually smooth time series.

The same investigation documents strategic intent and the organizational pathway of manipulation. According to the Ministry's notice, Zhongtan Energy Investment promised contractually that, if the allowance-allocation scheme and accounting method remained unchanged, it could turn a power plant's allowance deficit into a surplus. When enterprises had not retained historical coal samples, the agency instructed them to produce substitute samples for retrospective testing. Investigators also found multiple versions of production and coal-quality reports containing inconsistent values [14]. This evidentiary chain has high diagnostic value:

the apparent objective was to change the economic consequences of compliance, followed by retrospective reconstruction of source materials and alteration of metadata and accounting parameters. It strongly supports a mechanism combining reporting-time manipulation with record fabrication. Alternative explanations such as clerical error are less persuasive in light of the coordinated changes across dates, samples, identifiers, and numerical values. The contractual promise further connects the documentary alterations to the compliance outcome they were intended to change.

Three limitations qualify the implications of this case. First, the documented actors were regulated emitters and technical service agencies; the source does not establish involvement by Inner Mongolia provincial officials. Second, although the case involved enterprises in Inner Mongolia, the consulting agency was registered in Beijing and the Ministry's investigation was national. The case therefore cannot establish that Inner Mongolia as a whole is unusually prone to manipulation. Third, it shows neither that AI caused the falsification nor that AI-generated warnings uncovered it. The Ministry attributed the findings to targeted supervision and on-site inspection of technical service agencies and associated enterprises [14]. The incident instead identifies vulnerabilities that digital regulation must address and demonstrates the continuing importance of documentary and physical evidence obtained outside routine reporting channels. These limitations do not diminish the seriousness of the documented conduct. They delimit the level at which the inference is valid: a confirmed mechanism in an investigated organizational chain, not a province-wide prevalence estimate.

Later regulations addressed several of these vulnerabilities. The 2023 notice requires timely monthly archiving, documentary review, and on-site verification, with results recorded through the platform [2]. The *2024 Regulations* require retention of original records for five years and impose liability for data falsification [1]. These safeguards strengthen traceability and raise the cost of manipulation. Their practical effectiveness nevertheless remains difficult to verify because provincial materials do not disclose version logs, alert-confirmation rates, emissions corrections, or related penalties.

Evidence of platform capture is weaker still. The materials do not show a provincial authority, platform operator, or contractor altering model thresholds, suppressing alerts, deleting valid records, or denying raw data to higher-level reviewers. Zhejiang's presentation does not disclose revision permissions or model-validation results [11]; Hebei's response does not explain how daily data are generated or amended [12]; and Inner Mongolia's reports do not clarify how laboratory findings enter platform decisions [13]. These omissions indicate limited public auditability, not capture. Platform capture remains theoretically possible, but the available record is insufficient to establish it. An undisclosed authorization structure cannot itself be treated as evidence of improper interference. Demonstrating capture would require records of intervention in system design or operation, testimony about suppressed findings, discrepant logs, or another observable trace connecting an interested actor to a distorted regulatory output.

The defensible empirical conclusion is narrower than a three-mechanism account, but also more credible. Public evidence shows that regulated enterprises and intermediary agencies manipulated the timing and documentary basis of carbon-emissions reporting, and that central on-site inspection could identify that conduct. Composite indicators and opaque platform interfaces also create conditions that may invite strategic adaptation. The evidence does not establish indicator substitution or platform capture in the three provincial systems. This asymmetry supports the broader theoretical claim: digitization redistributes discretion without necessarily eliminating it. Regulatory credibility depends on whether digital records remain connected to independently verifiable physical evidence, external review, and regulatory consequences that can themselves be verified. The result also clarifies the relationship between reactivity and auditability. Better audit trails can

expose some manipulations after the fact, but actors may still adapt in ways that remain formally compliant and therefore require validity tests rather than conventional fraud detection.

4.3. From identification to enforcement closure: a cross-case comprehensive evaluation

Comparison across the three provinces yields a substantive but bounded answer to the research question (as shown in Table 1). AI-enabled regulation matters because it changes the state's capacity to acquire, organize, screen, and retain information. National and provincial records document reporting channels, model-assisted screening, electronic audit trails, composite indicators, and physical testing capacity, showing that digital tools are embedded in the regulatory process [2, 3, 11-13]. The evidence cannot attribute a particular penalty, report correction, or emissions change solely to AI. That boundary is central to the argument: technological capabilities operate mainly at the front end of regulation, whereas regulatory credibility emerges only when their outputs connect to independent verification, accountable decisions, and enforcement closure.

Table 1. National and provincial documents

Setting	Publicly documented contribution	Missing link	Warranted inference
National system	Monthly evidence retention, electronic verification records, rule- and model-based anomaly screening, AI-assisted information retrieval, and multilevel review [2, 3]	Case-level links among alerts, confirmed findings, corrections, sanctions, and follow-up	An operating digital MRV architecture expands screening and traceability; its regulatory value depends on review and closure
Zhejiang	Multi-source data integration, conversion models, sensors, and consequential carbon-efficiency codes [11]	Revision permissions, model validation, firm-level adaptation, and regulatory disposition	Integrated data are administratively actionable; credibility depends on indicator validity and auditability
Hebei	Coverage of more than 400 enterprises and a workflow of daily reporting, monthly archiving, and annual verification [12]	Reported variables, source-to-platform transmission, revision histories, and confirmed alerts	Frequent reporting and retention support longitudinal review; source integrity and enforcement use remain unclear
Inner Mongolia	Laboratory capacity for more than 50 parameters and testing of 74 coal-sample batches [13]	Sampling frame, discrepancies, corrected reports, sanctions, and links to platform decisions	Physical testing can supply evidence independent of enterprise calculations; its connection to regulatory disposition is undisclosed

The first comparative dimension is information integrity. National rules require monthly supporting materials to be archived, verification to leave electronic audit trails, and original records to be retained for at least five years [1, 2]. These requirements allow regulators to reconstruct what was reported, on what evidentiary basis, and when it was reviewed. The 2021 fraud case demonstrates why this matters: altered dates, removed QR codes, editable reports, and substitute coal samples made discrepancies across documents, versions, dates, and samples central to reconstructing the misconduct [14]. Documentation cannot guarantee authentic source data, but it can expose inconsistencies. Inner Mongolia's laboratories can also verify selected physical parameters independently of enterprise calculations [13]. Although the regulatory consequences of

those tests are undisclosed, the evidence supports a limited but concrete effect: digital traceability improves the auditability of documentation, while physical testing reduces exclusive reliance on enterprise self-reporting. Information integrity is therefore best understood as a property of the evidentiary chain, not of a single database. A secure platform cannot correct an unauthentic sample, but linked records can make inconsistencies discoverable and assign responsibility for later changes.

The second dimension is measurement validity. Zhejiang's Carbon Efficiency Code converts integrated data into consequential administrative indicators used for enterprise comparison and resource allocation [11]. Digitization thus standardizes information dispersed across agencies and makes it actionable. Precisely because the indicators carry consequences, however, their definitions and weights require scrutiny. Public sources do not establish whether higher scores reliably represent broader policy objectives or whether enterprises adapt to indicator boundaries. Hebei's briefing does not identify the variables included in daily reports [12], so reporting frequency cannot be treated as evidence of emissions accuracy. Inner Mongolia's coal-sample testing can verify selected physical parameters but cannot establish the representativeness of an entire emissions-accounting system. AI and digital tools can direct regulatory attention and enable comparison; measurement validity still depends on indicator design, data coverage, and external validation. Validity also has both technical and institutional components. A statistically stable model may still privilege an administratively convenient proxy, while a conceptually appropriate indicator may become misleading when important activities fall outside its coverage.

The third dimension is enforcement closure. A 2023 national notice establishes a formal sequence: technical reviewers identify anomalies, municipal authorities investigate and require rectification, and verified findings return through the national platform [2]. The 2025 national report states that more than three million power-sector records were reviewed, problematic data fell by approximately 30 percent from the previous year, noncompliant items in verification reports fell by about 24 percent, and five unannounced inspections targeted anomalous composite indicators or inadequate rectification [3]. These figures show an operating review and follow-up system rather than a merely symbolic deployment. They do not reveal which warning prompted which investigation, however, and the changes cannot be attributed solely to AI. Likewise, the 2022 fraud bulletin states that investigations would continue but does not report final dispositions for the Inner Mongolia enterprises [14]. The institutional path from screening to intervention is therefore observable in aggregate, while closure in individual cases remains unavailable for independent external verification. Closure requires more than evidence that an agency received a referral. It requires a documented disposition, an identified responsible actor, and verification that any ordered correction was completed and persisted beyond the immediate review.

The appropriate unit of analysis is AI-enabled regulation as a sociotechnical architecture, not an isolated algorithmic treatment. The national report describes AI-assisted information retrieval and anomaly screening that combines rules, models, parameter relationships, and knowledge graphs [3]. These tools operate alongside digital reporting, administrative task assignment, and human review. The analytical question is how this hybrid arrangement redistributes capacity for observation and screening, and under what conditions its outputs acquire regulatory authority [15, 16]. The provincial cases show that technology has entered collection, screening, and verification and expose important institutional bottlenecks, but they cannot identify AI's independent contribution to final outcomes.

The evidence provides differentiated support for the three propositions. Proposition 1 receives strong support at the level of institutional capacity: platform access, record retention, model-assisted screening, and physical testing reduce reporting entities' exclusive control over information [1-3, 11-13]. Their case-level effectiveness cannot yet be quantified, but the resulting information structure is materially different from

unverified enterprise self-reporting. Proposition 2 receives direct behavioral support from reporting-time manipulation and record fabrication [14]. Indicator substitution remains a documented institutional vulnerability rather than an observed violation, and platform capture lacks evidentiary support. Proposition 3 is supported by formal design and aggregate national practice: anomalies can be assigned for investigation and rectification, and national records document review and follow-up [2, 3]. Yet closure in individual cases remains independently unauditable. This imbalance is itself a finding: publicly observable digitalization is substantially further developed in information acquisition and screening than in final disposition. The propositions therefore receive support at different levels of analysis: institutional capacity, documented behavior, and aggregate administrative practice. Stating those levels explicitly avoids presenting partial confirmation as a uniform test of the full causal chain.

The cross-case conclusion is therefore conditional. AI-enabled tools reshape the first stage of the principal-agent relationship by expanding the volume, frequency, standardization, searchability, and traceability of information available to regulators. Their contribution to credible regulation depends on four linkages: access to evidence not controlled by reporting entities; traceable revision histories; valid connections between metrics and policy objectives; and documented, verifiable resolution of identified anomalies. Where these linkages exist, technological visibility can support actionable oversight. Where they are absent or externally opaque, the same technology may produce procedural visibility without demonstrable accountability. The evidence does not justify classifying any province as a digital facade, nor does deployment itself establish success. Instead, the analysis identifies the institutional threshold between a useful digital tool and a digital facade. The recommendations below focus on strengthening and disclosing the links from data generation to final disposition. The four linkages are cumulative rather than interchangeable. Strong source evidence cannot rescue an invalid policy metric, and accurate detection cannot compensate for an enforcement process that leaves corrective action undocumented.

5. Recommendation

The findings indicate that reform should strengthen the institutions surrounding AI rather than focus primarily on further technology procurement. AI-enabled tools have expanded information processing and screening. The remaining challenge is to make the path from source data to final disposition independently auditable while preserving necessary human judgment and assigning clear responsibility for its exercise.

First, establish auditable data provenance and version control. The Ministry of Ecology and Environment could adopt uniform standards linking each monthly submission to its source documents, instrument readings, fuel samples, calculations, and responsible entities. Every revision should preserve the original value and record who made the change, when, and why. Supervising reviewers should receive read-only access to original evidence rather than relying solely on provincial summaries or the latest submitted version. Existing requirements for electronic audit trails and five-year retention of original records provide an institutional foundation [1, 2]. The 2021 fraud cases show why dates, report versions, QR codes, and sample identifiers must remain retrospectively reconstructable [14]. For calculated emissions, regulators could use risk-based random sampling and require physical samples to be tested by organizations independent of reporting enterprises and their consultants. Inner Mongolia's laboratory capacity provides a possible institutional base for this function [13]. Legitimate corrections should remain possible without overwriting the original evidentiary chain. A common provenance standard would also make records portable across administrative levels and technical providers. That portability matters when a reviewer must reconstruct a case after staff, software, or contractors have changed.

Second, separate algorithmic detection from independent verification. Models, composite indicators, and anomaly rules should prioritize audits rather than serve as conclusive evidence of compliance. Regulators should document each tool's purpose, inputs, validation procedure, update cycle, and known error conditions. Such disclosure is especially important when indicators affect ratings, credit, or resource allocation, as in Zhejiang, or when high-frequency reporting produces large screening volumes, as in Hebei [11, 12]. Reviewers organizationally independent of reporting entities and platform providers should examine both flagged and unflagged cases against source documents and physical evidence. Sampling unflagged cases can reveal adaptation around known thresholds. Human reviewers should record why alerts were confirmed or dismissed, what evidence informed the decision, and what action followed. This arrangement preserves the speed of automated screening without allowing opaque scores to become self-validating regulatory judgments. Organizational separation need not require a wholly separate agency in every case. It does require that the verifier have access to original evidence, freedom to disagree with the model output, and responsibility to document the basis of that judgment.

Third, publish standardized indicators of enforcement closure. For each reporting cycle, national and provincial reports should disclose the number of alerts generated, cases manually reviewed, alerts confirmed and dismissed, emissions reports corrected, matters referred for penalties or rectification, and corrective actions verified as complete. Anonymous case identifiers could connect these stages without disclosing confidential business information or model code. Reports should also give median resolution times and post-rectification recurrence rates. National totals for reviews and declining issue rates show that regulatory activity occurs [3], but they do not establish whether a particular signal produced a sustained correction. Standardized closure reporting would allow external evaluation of the regulatory chain, avoid treating nondisclosure as proof of failure, and prevent penalty counts from becoming the sole measure of effectiveness.

Together, these reforms would convert numerical visibility into verifiable accountability. Human discretion remains necessary; the objective is to make its evidentiary basis, exercise, and consequences traceable from initial observation to final resolution.

6. Conclusion

This paper examines when AI-enabled monitoring can strengthen the credibility of China's multilevel carbon regulation and when it risks becoming a digital facade. The three provincial cases support a conditional answer. Digital tools can connect dispersed data, increase reporting frequency, standardize comparison, preserve records, and help regulators prioritize anomalies. Deployment alone, however, neither verifies source data nor ensures that an alert leads to corrective action. AI changes the information architecture of regulation. That change produces credible regulation only when technical outputs connect to independently accessible evidence, traceable revisions, valid metrics, and documented final dispositions. The decisive issue is thus not whether the regulatory system appears technologically advanced, but whether its institutional design permits independent reconstruction of how a digital signal became, or failed to become, a regulatory outcome.

Public evidence is strongest for information acquisition and screening. The national system establishes record-retention and review pathways; Zhejiang shows how digital indicators can influence administrative and financial decisions; Hebei illustrates recurring reporting and verification; and Inner Mongolia supplies direct evidence of manipulation involving dates, documents, and sample identities alongside newer physical-testing capacity [1, 2, 11-14]. Direct evidence therefore supports reporting-time manipulation. Indicator substitution is a documented institutional vulnerability, while platform capture remains unsubstantiated. National aggregate materials show an operating review system but do not connect algorithmic flags to sustained corrective action

[3]. The paper consequently neither ranks provincial credibility nor labels any province a digital facade. It identifies the institutional threshold separating effective digital tools from digital facades. These findings also explain why differences in public visibility should not be converted into a provincial league table. The cases disclose different parts of the same regulatory chain, and the absence of comparable case-level closure data prevents a defensible ordinal ranking.

The paper's theoretical contribution is to conceptualize AI-enabled regulation as a sociotechnical institution rather than an autonomous technological intervention. AI can ease information bottlenecks without eliminating discretion, which may migrate to data selection, indicator construction, validation, and exception handling. Credible regulation therefore requires auditability across the regulatory chain. Human judgment remains indispensable, but both its evidentiary basis and its consequences must be reviewable. This perspective also places limits on technological solutionism. Improving models may increase sensitivity or efficiency, but institutional credibility depends on who can inspect the evidence, challenge an output, authorize action, and verify completion.

Public documents do not permit estimation of AI's independent causal effect or reveal complete platform logs. This limits the paper's inferences without negating the institutional changes that can be observed. Future research should use anonymous case identifiers to connect warnings, timestamps, version histories, investigations, corrective actions, and follow-up outcomes and, where possible, combine those records with interviews and inspection files. The policy implication is clear: technology becomes a credible regulatory instrument only when digital visibility is converted into verifiable accountability. Otherwise, technical precision and procedural activity may coexist without demonstrable enforcement closure. A stronger empirical design would permit researchers to distinguish false positives, corrected errors, deliberate manipulation, and unresolved anomalies. It would also allow assessment of whether improvements persist after the immediate pressure of inspection has passed.

References

[1] Tang, R., Guo, W., Oudenes, M., Li, P., Wang, J., Tang, J., Wang, L., & Wang, H. (2018). Key challenges for the establishment of the monitoring, reporting and verification (MRV) system in China's national carbon emissions trading market. *Climate Policy, 18*(sup1), 106–121. https://doi.org/10.1080/14693062.2018.1454882

[2] Zeng, X., Duan, M., Yu, Z., Li, W., Li, M., & Liang, X. (2018). Data-related challenges and solutions in building China's national carbon emissions trading scheme. *Climate Policy, 18*(sup1), 90–105. https://doi.org/10.1080/14693062.2018.1473239

[3] Yu, R., Zhang, D., Zhang, X., & Huang, X. (2023). Machine learning for data verification in emissions trading system. *Resources, Conservation and Recycling, 199*, 107239. https://doi.org/10.1016/j.resconrec.2023.107239

[4] Greenstone, M., He, G., Jia, R., & Liu, T. (2022). Can technology solve the principal-agent problem? Evidence from China's war on air pollution. *American Economic Review: Insights, 4*(1), 54–70.

[5] Espeland, W. N., & Sauder, M. (2007). Rankings and reactivity: How public measures recreate social worlds. *American Journal of Sociology, 113*(1), 1–40.

[6] Yang, L., Lin, Y., Wang, J., & Peng, F. (2024). Achieving air pollution control targets with technology-aided monitoring: Better enforcement or localized efforts? *American Economic Journal: Economic Policy, 16*(4), 280–315.

[7] O'Brien, K. J., & Li, L. (1999). Selective policy implementation in rural China. *Comparative Politics, 31*(2), 167–186.

[8] Mertha, A. (2009). "Fragmented authoritarianism 2.0": Political pluralization in the Chinese policy process. *The China Quarterly, 200*, 995–1012.

[9] Kostka, G. (2016). Command without control: The case of China's environmental target system. *Regulation & Governance, 10*(1), 58–74.

[10] van Noordt, C., & Misuraca, G. (2022). Artificial intelligence for the public sector: Results of landscaping the use of AI in government across the European Union. *Government Information Quarterly, 39*, 101714.

[11] Li, R., Lu, F., Xu, J., Chen, K., & Zhao, X. (2023). Effect of carbon information disclosure with consistent evaluation standards: An empirical study about carbon efficiency label in Huzhou China. *Energy Policy, 181*, 113717. https://doi.org/10.1016/j.enpol.2023.113717

[12] Li, Y., Li, M., Hou, C., Gao, Y., Wang, Q., & Tian, J. (2024). Construction and implementation of a digital carbon management system for carbon neutrality of steel industry. *Iron and Steel, 59*(9), 13–21. https://doi.org/10.13228/j.boyuan.issn0449-749x.20240342

[13] Hu, Z. (2023). Thinking on improving carbon emission data quality of coal power enterprises by measurement. *Industrial Metrology, 33*(5), 75–79. https://doi.org/10.13228/j.boyuan.issn1002-1183.2022.0248

[14] Wu, Q., Smits, M., & Zhu, A. L. (2024). Strategic transparency under authoritarian environmentalism: Information disclosure and the role of environmental NGOs in China's national emission trading scheme. *Climate Policy*, 1–20. https://doi.org/10.1080/14693062.2024.2386380

[15] Yeung, K. (2018). "Algorithmic regulation": A critical interrogation. *Regulation & Governance, 12*(4), 505–523.

[16] Danaher, J., Hogan, M. J., Noone, C., Kennedy, R., Behan, A., De Paor, A., Felzmann, H., Haklay, M., Khoo, S.-M., Morison, J., Murphy, M. H., O'Brolcháin, N., Schafer, B., & Shankar, K. (2017). Algorithmic governance: Developing a research agenda through the power of collective intelligence. *Big Data & Society, 4*(2), 1–21.

[17] Kostka, G., & Hobbs, W. (2012). Local energy efficiency policy implementation in China: Bridging the gap between national priorities and local interests. *The China Quarterly, 211*, 765–785.

[18] van der Kamp, D., Lorentzen, P., & Mattingly, D. (2017). Racing to the bottom or to the top? Decentralization, revenue pressures, and governance reform in China. *World Development, 95*, 164–176.

[19] George, A. L., & Bennett, A. (2005). *Case studies and theory development in the social sciences*. MIT Press.

[20] Seawright, J., & Gerring, J. (2008). Case selection techniques in case study research: A menu of qualitative and quantitative options. *Political Research Quarterly, 61*(2), 294–308.

[21] Collier, D. (2011). Understanding process tracing. *PS: Political Science & Politics, 44*(4), 823–830.