WORDS TO COORDINATES: AN UNCERTAINTY-AWARE NLP–GEOSPATIAL FRAMEWORK FOR MEASURING THE GAP BETWEEN DISCLOSED AND PHYSICAL CLIMATE RISK

A PREPRINT

Thanatchon Chongmankhong

November 2026

ABSTRACT

Corporate climate disclosure is being standardised rapidly, with IFRS S2 emerging as a global baseline, yet the information that matters most for physical climate risk—where a firm's assets are and which hazards they face—remains embedded in unstructured, multilingual reports. Text-based measures of firm climate exposure capture what firms say but not what their assets face, while geospatial assessments rely on curated asset databases and ignore disclosure content. We propose *Words to Coordinates* (W2C), a framework that links the two. W2C extracts taxonomy-aligned climate disclosures and asset locations from reports with page-level provenance, resolves assets to coordinates, overlays them with scenario-based hazard projections, and compares modelled exposure with disclosed risk through a Disclosure Gap Index that places firms in four quadrants: aware, silent risk, cheap talk and benign. Two design choices distinguish W2C: distant supervision from XBRL-tagged filings, to extend coverage to under-served and non-English markets, and end-to-end propagation of extraction, geolocation and hazard-model uncertainty, so that every classification carries a calibrated probability. We derive a closed-form relation between geolocation error and the fidelity of exposure estimates, which implies that location error should not exceed one third of a hazard's spatial correlation length. A simulation study on the real locations of 11,107 power plants held by 1,247 owners in 76 countries confirms the relation (mean absolute deviation 0.002) and shows that (i) country-level geocoding preserves a firm-level rank correlation of at most 0.54, even for spatially coherent hazards; (ii) for local hazards, city-level geolocation error degrades quadrant classification more than five times as much as a 20% extraction error rate; and (iii) full uncertainty propagation reduces expected calibration error from 0.075 to 0.033 and supports selective identification of silent-risk firms at 84% precision. We close with an evaluation protocol and identification strategies for a global empirical application.

Keywords Climate risk disclosure · Physical climate risk · Natural language processing · Geospatial analysis · Uncertainty quantification · Sustainable finance

1 Introduction

Climate-related financial disclosure is moving from voluntary frameworks to mandatory, standardised reporting. Building on the recommendations of the Task Force on Climate-related Financial Disclosures (Task Force on Climate-related Financial Disclosures, 2017), the International Sustainability Standards Board (ISSB) issued IFRS S1 and IFRS S2 in June 2023, effective for annual reporting periods beginning on or after 1 January 2024 (International Sustainability Standards Board, 2023a,b). A growing number of jurisdictions are adopting or adapting these standards, alongside the European Sustainability Reporting Standards (ESRS) (European Commission, 2023). Institutional investors report that climate risks matter for their decisions (Krueger et al., 2020), yet the data on which those decisions rest are contested: ESG ratings from different providers diverge substantially (Berg et al., 2022), and so do commercial metrics of physical climate risk (Hain et al., 2022).

Two research traditions measure firm-level climate risk from different raw material. The first is textual. Climate exposure has been measured from earnings-call transcripts (Sautner et al., 2023), and domain-adapted language models such as ClimateBERT (Webersinke et al., 2021) have been used to classify climate disclosures, detect environmental claims and quantify the specificity of corporate commitments (Bingler et al., 2022; Stammbach et al., 2023; Schimanski et al., 2024a). Large language models now support question answering over sustainability reports (Ni et al., 2023) and are evaluated on retrieval benchmarks built from corporate climate disclosures (Schimanski et al., 2024b). These measures capture what firms *say*, but not what their physical assets *face*. The second tradition is geospatial. Spatial finance overlays the locations of physical assets with hazard data to estimate exposure from the bottom up (Caldecott et al., 2022), using projections derived from climate model ensembles (Eyring et al., 2016; Thrasher et al., 2022) and hazard maps such as Aqueduct (Ward et al., 2020; Kuzma et al., 2023). These measures capture what assets face, but they rely on curated asset databases that cover a limited set of sectors and markets, and they are silent about what firms disclose.

The question that connects the two traditions is, in our view, among the most policy-relevant in sustainable finance: *do firms disclose the physical climate risks to which their assets are actually exposed, and where are the gaps largest?* Supervisors need the answer to target reviews, investors to price undisclosed risk, and assurance providers to test the completeness of disclosures. Answering it requires linking words to coordinates at scale: extracting both disclosure content and asset locations from the same documents, across languages and reporting regimes, with traceable provenance and quantified uncertainty. Coverage compounds the difficulty. ESG data are concentrated on large firms in developed markets (Kotsantonis and Serafeim, 2019), leaving *ESG data deserts* in emerging markets, where physical hazards are often severe and disclosures appear in local languages and unstructured formats.

This paper makes four contributions.

1. **Measurement.** We formalise the *Disclosure Gap Index* (DGI), which compares a firm's modelled physical exposure with its disclosure intensity within a peer group, and the *disclosure–exposure plane*, which classifies firms as aware, silent risk, cheap talk or benign (Section 3).
2. **Framework.** We specify Words to Coordinates (W2C), an end-to-end architecture that is provenance-first (every extracted value is linked to a document, page and bounding box), taxonomy-aligned (an IFRS S2 schema with crosswalks to ESRS E1 and GRI 305) and coverage-oriented, using XBRL-tagged filings as distant supervision for multilingual extraction (Section 4).
3. **Theory.** We derive a closed-form relation between geolocation error and the correlation between true and estimated hazard exposure. It yields a simple design rule—location error should not exceed one third of the hazard's spatial correlation length—that tells a pipeline how precise its geocoding must be for each hazard (Section 3.3).
4. **Evidence on measurement properties.** A simulation study grounded in the real locations of 11,107 power plants confirms the theory, quantifies how extraction and geolocation errors combine in quadrant classification, and shows that end-to-end uncertainty propagation yields calibrated probabilities that support selective identification of silent-risk firms (Section 5).

Scope. This is a framework and methods paper. It does not report disclosure gaps for real firms: the disclosure variables and hazard fields in our simulation are synthetic, and only the asset geography is real. We instead establish the measurement properties of the DGI under controlled conditions in which the ground truth is known, and set out the evaluation protocol (Section 6) and identification strategies (Section 7) for the empirical programme that the framework is designed to support. Section 8 discusses limitations and ethical considerations, and Section 9 concludes.

2 Related work

NLP for climate and ESG disclosure. Early work established datasets for detecting climate-related text (Varini et al., 2020) and verifying climate claims (Diggelmann et al., 2020), and applied NLP to sustainability reports (Luccioni et al., 2020). ClimateBERT (Webersinke et al., 2021) enabled analyses showing that much corporate climate disclosure is non-specific "cheap talk" (Bingler et al., 2022) and that regulatory disclosure of transition and physical risks is reflected in credit default swap spreads (Kölbel et al., 2024). Subsequent work extended measurement to environmental claims (Stammbach et al., 2023) and to ESG communication more broadly (Schimanski et al., 2024a), and used retrieval-augmented generation (Lewis et al., 2020) for report analysis (Ni et al., 2023; Schimanski et al., 2024b). Two gaps motivate W2C. First, these systems output labels or answers rather than structured, taxonomy-aligned values with provenance and calibrated uncertainty. Second, they rarely extract the geographic information—sites, facilities and locations—that would allow disclosures to be checked against physical exposure.

Physical climate risk and spatial finance. Physical risk assessment combines hazard, exposure and vulnerability. Hazard projections derive from CMIP6 ensembles (Eyring et al., 2016) under ScenarioMIP pathways (O'Neill et al., 2016), often statistically downscaled (Thrasher et al., 2022); specialised products cover floods and water stress (Ward et al., 2020; Kuzma et al., 2023), tropical cyclones (Knapp et al., 2010) and supervisory scenarios (Network for Greening the Financial System, 2023). Spatial finance links such hazard data to asset-level datasets (Caldecott et al., 2022), including open registries of power plants (Byers et al., 2018), energy infrastructure (Global Energy Monitor, 2024) and emission sources (Climate TRACE, 2024). Hain et al. (2022) show that physical risk scores from different providers disagree substantially. Text-based measures (Sautner et al., 2023) sidestep the need for locations, but measure attention to climate topics rather than exposure. W2C treats location as an uncertain, extracted quantity and propagates its uncertainty.

Disclosure standards as an extraction target. IFRS S2 requires disclosure of climate-related risks and opportunities (¶ 10), of climate resilience assessed with scenario analysis (¶ 22), of cross-industry metrics including Scope 1–3 greenhouse gas emissions and the amount and percentage of assets or business activities vulnerable to physical risks (¶ 29), and of climate-related targets (¶¶ 33–36) (International Sustainability Standards Board, 2023b). ESRS E1 (European Commission, 2023) and GRI 305 (Global Reporting Initiative, 2016) contain overlapping requirements, and the TNFD recommendations make location the explicit first step of nature-related assessment (Taskforce on Nature-related Financial Disclosures, 2023). Standardisation turns disclosures into a well-defined extraction target, and digital taxonomies with XBRL tagging, where they exist, provide machine-readable labels that W2C exploits as distant supervision.

Distant supervision, calibration and uncertainty. Distant supervision aligns an existing knowledge base with text to create noisy training labels (Mintz et al., 2009), and knowledge distillation transfers the behaviour of a large model to a smaller one (Hinton et al., 2015). Probabilistic predictions should be calibrated (Guo et al., 2017) and evaluated with proper scoring rules (Brier, 1950); conformal prediction provides distribution-free coverage guarantees for set-valued outputs (Vovk et al., 2005; Angelopoulos and Bates, 2023), and selective classification formalises abstention (Geifman and El-Yaniv, 2017). Measurement error in explanatory variables biases downstream estimates (Carroll et al., 2006), which is why W2C carries uncertainty through to its outputs rather than discarding it after extraction.

3 Problem formulation

3.1 Setup

Consider a set of firms $\mathscr{I}$ in a comparison set, for example a sector–year. Firm i controls assets $a \in \mathscr{A}_i$ with locations $\mathbf{x}_a$ on the sphere and non-negative weights w_a, $\sum_{a\in\mathscr{A}_i} w_a = 1$, that reflect the share of capacity, asset value or revenue attributable to each asset. Physical hazards $h \in \mathscr{H}$ (riverine and coastal flooding, extreme heat, drought, tropical cyclones, wildfire) are described by hazard fields $Z_{h,s,t}(\mathbf{x})$ under scenario s and time horizon t, available as an ensemble $\{Z^{(m)}_{h,s,t}\}_{m=1}^{M}$ of model outputs. We suppress h,s,t when unambiguous.

Definition 1 (Modelled exposure)**.** The modelled exposure of firm i is

$$E_i = \sum_{a\in\mathscr{A}_i} w_a\, g\big(Z(\mathbf{x}_a)\big), \tag{1}$$

where g is a monotone transform mapping hazard intensity to a normalised exposure or damage score.

Definition 2 (Disclosure intensity)**.** For each hazard, the disclosure intensity $D_i \in \{0,\dots,4\}$ of firm i is an ordinal level:

0. *None*: no reference to the physical risk;
1. *Generic*: physical climate risk mentioned without hazard specificity;
2. *Hazard-specific*: the hazard is identified and linked to operations or locations;
3. *Quantified*: exposure or anticipated financial effects are quantified (cf. IFRS S2 ¶ 29(c));
4. *Quantified with scenarios and adaptation*: quantified effects under named scenarios, with a resilience or adaptation response (cf. IFRS S2 ¶ 22).

The scale is hazard-specific: a firm may disclose flood risk at level 3 and heat risk at level 0.

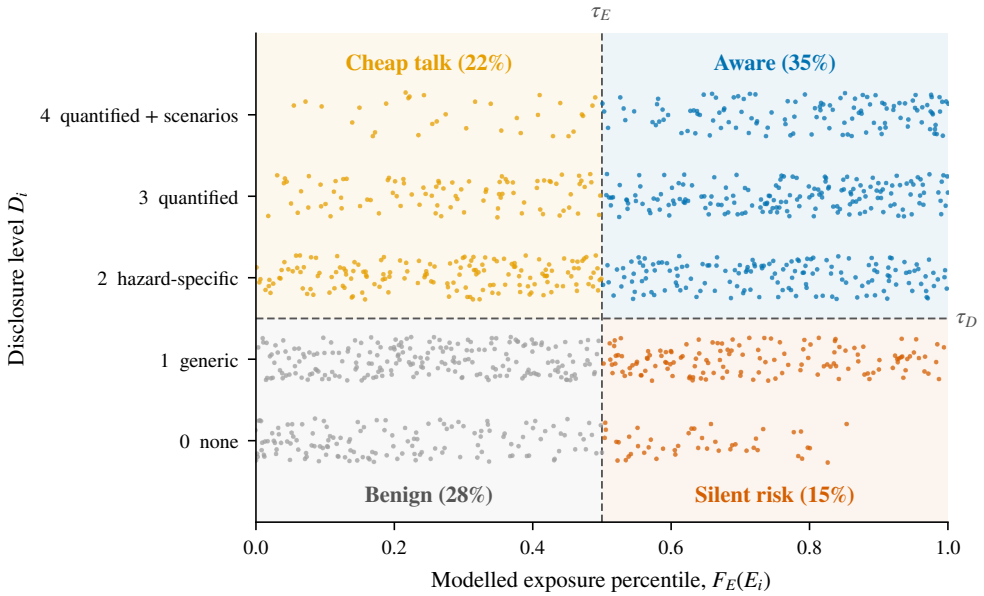


Figure 1. The disclosure–exposure plane for one simulated replicate (1,247 firms; real asset locations, synthetic hazard with $\ell = 50$ km and synthetic disclosures). Points are jittered vertically. Dashed lines mark the thresholds $\tau_E = 0.5$ and $\tau_D = 2$; percentages give the share of firms in each quadrant.

Definition 3 (Disclosure Gap Index)**.** Let F_E and F_D denote the mid-rank empirical distribution functions of E and D within the comparison set. The Disclosure Gap Index of firm i is

$$\mathrm{DGI}_i = F_E(E_i) - F_D(D_i) \in (-1, 1). \tag{2}$$

Positive values indicate disclosure below what the firm's relative exposure would suggest; negative values indicate disclosure beyond it. Because both terms are peer-relative, the DGI is invariant to monotone transformations of E and comparable across hazards and sectors.

Definition 4 (Disclosure–exposure quadrants)**.** Given thresholds $\tau_E \in (0,1)$ and $\tau_D \in \{1,\dots,4\}$, firm i is classified as

$$Q_i = \begin{cases} \text{aware} & \text{if } F_E(E_i) > \tau_E \text{ and } D_i \geq \tau_D, \\ \text{silent risk} & \text{if } F_E(E_i) > \tau_E \text{ and } D_i < \tau_D, \\ \text{cheap talk} & \text{if } F_E(E_i) \leq \tau_E \text{ and } D_i \geq \tau_D, \\ \text{benign} & \text{otherwise.} \end{cases} \tag{3}$$

We use $\tau_E = 0.5$ and $\tau_D = 2$ throughout, so that *silent risk* denotes an above-median exposed firm that discloses no hazard-specific information. Figure 1 illustrates the plane. We borrow the term *cheap talk* from Bingler et al. (2022) to denote disclosure intensity that exceeds modelled exposure; the label is descriptive rather than normative, because such disclosure may reflect exposures outside the hazard model, such as supply-chain risks, or prudent reporting.

3.2 Uncertainty and probabilistic classification

None of the inputs to (1)–(3) is observed without error. A pipeline observes (i) an extracted level $\hat{D}_i$ with a calibrated predictive distribution $P(D_i \mid \hat{D}_i)$; (ii) estimated locations $\hat{\mathbf{x}}_a$ with a precision class $r(a)$ and an error scale $\sigma_{r(a)}$; and (iii) an ensemble of hazard models rather than the hazard itself. We therefore treat DGI_i and Q_i as random variables and report $P(Q_i = q \mid \text{data})$ together with quantiles of DGI_i, computed by Monte Carlo (Algorithm 2). Because exposure enters through peer-relative ranks, the classification of one firm depends on the uncertain exposures of all firms; the Monte Carlo therefore recomputes ranks in every draw. A firm is flagged as silent risk only if $P(Q_i = \text{silent risk}) \geq \tau_p$; otherwise the pipeline abstains, which is a selective classification rule (Geifman and El-Yaniv, 2017).

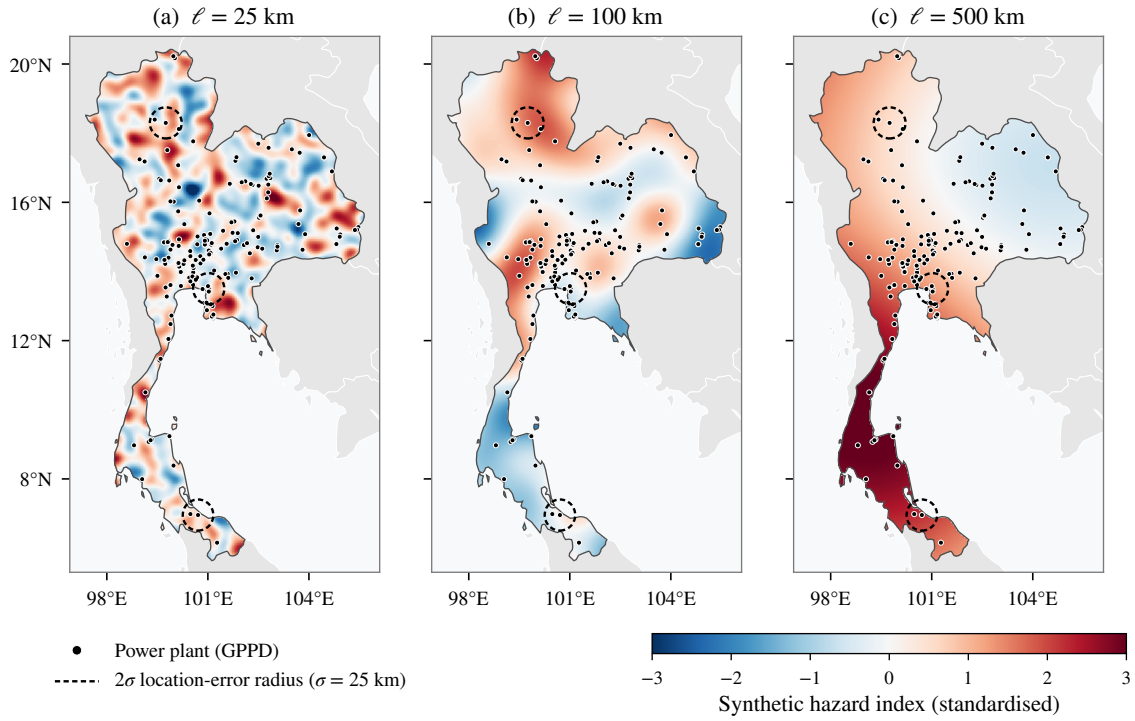


Figure 2. The same synthetic hazard pattern rendered at three correlation lengths over Thailand, with the 196 Thai plants in the Global Power Plant Database (Byers et al., 2018). Dashed circles show a 2σ radius for city-level geolocation error ($\sigma = 25$ km) around three large plants. For a local hazard (a), the circle spans several independent hazard features, so an exposure estimate from the wrong location is close to a random draw; for a regional hazard (c), the hazard is nearly constant within the circle. The fields are synthetic and do not represent actual hazards in Thailand.

3.3 How precise must geolocation be?

Geolocation is the step most often skipped or approximated: firms are frequently assigned to the country or city of their headquarters. The precision required depends on the hazard. Model a hazard locally as a stationary Gaussian random field with unit variance and squared-exponential correlation $\rho(d) = \exp\{-d^2/(2\ell^2)\}$, where ℓ is the correlation length, and model geolocation error as isotropic Gaussian, $\hat{\mathbf{x}} = \mathbf{x} + \boldsymbol{\delta}$ with $\boldsymbol{\delta} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_2)$ independent of the field.

Proposition 1. *Under these assumptions,*

$$\operatorname{Corr}\big(Z(\mathbf{x}), Z(\hat{\mathbf{x}})\big) = \frac{\ell^2}{\ell^2 + \sigma^2}. \tag{4}$$

Moreover, for any strictly increasing g, the Spearman correlation between $g(Z(\mathbf{x}_a))$ and $g(Z(\hat{\mathbf{x}}_a))$ across assets that are mutually independent (separated by distances large relative to ℓ) is

$$\rho_S(\sigma/\ell) = \frac{6}{\pi} \int_0^\infty \arcsin\left(\tfrac{1}{2}\, e^{-(\sigma/\ell)^2 t}\right) e^{-t}\, dt. \tag{5}$$

Both quantities depend on σ and ℓ only through the ratio σ/ℓ.

The proof is given in Appendix A. Inverting (4) gives a design rule.

Corollary 1 (Design rule)**.** $\operatorname{Corr}(Z(\mathbf{x}), Z(\hat{\mathbf{x}})) \geq \rho^*$ *if and only if* $\sigma \leq \ell\sqrt{1/\rho^* - 1}$. *In particular, preserving a correlation of* 0.9 *requires* $\sigma \leq \ell/3$.

The rule turns an abstract accuracy target into a per-hazard engineering specification. As a stylised guide, hazards such as riverine flooding vary over distances of a few kilometres, whereas heat extremes are typically coherent over hundreds of kilometres; the corresponding precision requirements differ by two orders of magnitude. Figure 2 illustrates the intuition with real plant locations in Thailand.

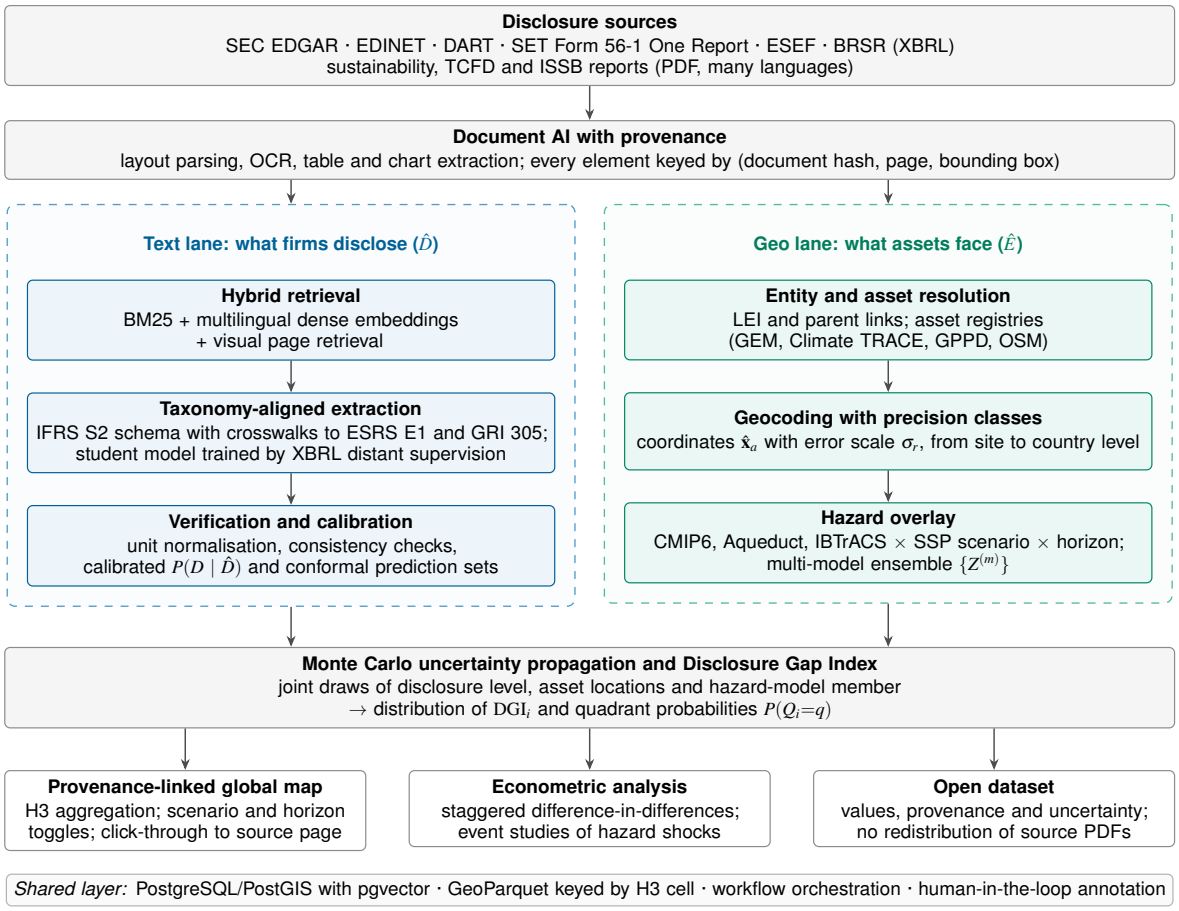


Figure 3. The W2C architecture. Documents are parsed once with provenance and then processed in a text lane, which estimates disclosure intensity $\hat{D}$ with a calibrated predictive distribution, and a geo lane, which estimates modelled exposure $\hat{E}$ from resolved, geocoded assets and a hazard-model ensemble. The two lanes are reconciled by Monte Carlo propagation (Algorithm 2).

Remark 1 (Finite domain)**.** Proposition 1 describes independent assets. When assets are clustered relative to ℓ, correlations computed within a single sample fall slightly below (4), because the sample variance of exposure shrinks. With $\bar{c}$ the mean pairwise correlation $\rho(d_{ab})$ among assets, the within-sample analogue of (4) is approximately $(\rho - \bar{c})/(1-\bar{c})$, where $\rho = \ell^2/(\ell^2+\sigma^2)$ (Appendix A).

4 The Words to Coordinates framework

Figure 3 summarises the architecture. Its organising idea is that disclosure (what firms say) and exposure (what assets face) are extracted from the same documents but processed in two lanes with different error models, and are reconciled only at the end, through the Monte Carlo step that produces the DGI and quadrant probabilities.

4.1 Ingestion across jurisdictions

W2C ingests periodic filings and voluntary reports from regulatory repositories and company websites (Table 1). Machine-readability differs sharply across regimes: some tag sustainability data in XBRL, while others publish designed PDF reports. W2C stores a cryptographic hash and retrieval metadata for every document, so that extracted data can be published without redistributing source files.

4.2 Document AI with provenance

Sustainability reports are design-heavy: key metrics often appear in tables, charts and infographics rather than in running text. W2C parses documents with layout-aware conversion (Auer et al., 2024), applies optical character recognition

Table 1. Disclosure sources and their machine-readability. Status as of mid-2026; regimes are evolving and should be re-verified at the time of data collection.

Jurisdiction	Repository and document	Language(s)	Machine-readable structure
United States	SEC EDGAR: Form 10-K	English	Financial statements in Inline XBRL; climate narrative untagged
European Union	ESEF annual financial reports; ESAP (planned)	Multiple	Financial statements in Inline XBRL; ESRS digital tagging forthcoming
Japan	EDINET: annual securities reports	Japanese	XBRL; sustainability section largely tagged as text blocks
South Korea	DART	Korean	XBRL for financial statements
India	NSE and BSE: BRSR	English	XBRL, including numeric ESG indicators
Thailand	SET: Form 56-1 One Report	Thai, English	Predominantly PDF
Global	Sustainability, TCFD and ISSB reports	Many	Predominantly designed PDF

Table 2. Core fields of the W2C extraction schema, with indicative locations of the corresponding requirements in IFRS S2 (International Sustainability Standards Board, 2023b), the ESRS (European Commission, 2023) and GRI 305 (Global Reporting Initiative, 2016). "Derived" fields are constructed by W2C rather than required verbatim. ESRS references follow the 2023 delegated act; later revisions may renumber requirements.

Schema field	IFRS S2	ESRS	GRI
Scope 1 GHG emissions	¶ 29(a)	E1-6	305-1
Scope 2 GHG emissions, location- and market-based	¶ 29(a)	E1-6	305-2
Scope 3 GHG emissions by category	¶ 29(a)	E1-6	305-3
Emission-reduction targets (base year, target year, coverage)	¶¶ 33–36	E1-4	–
Physical risks identified, by hazard and time horizon	¶ 10	ESRS 2 IRO-1	–
Climate resilience and scenario analysis	¶ 22	ESRS 2 SBM-3	–
Assets or business activities vulnerable to physical risks	¶ 29(c)	E1-9	–
Internal carbon price	¶ 29(f)	E1-8	–
Climate-related capital deployment	¶ 29(e)	E1-3	–
Disclosure level D per hazard (Definition 2)	derived	derived	–
Asset and site mentions (name, address, coordinates)	derived	derived	–

where needed, and keys every element—paragraph, table cell or chart—by (document hash, page, bounding box). Provenance is not a convenience. It allows each extracted value to be audited, it enables the map's click-through from an asset to the passage that mentions it, and it aligns the pipeline with the assurance regimes being introduced for sustainability reporting.

4.3 Taxonomy-aligned extraction

Retrieval combines lexical BM25 (Robertson and Zaragoza, 2009), multilingual dense embeddings (Chen et al., 2024) and visual page retrieval for chart- and table-heavy pages (Faysse et al., 2024). Extraction targets a typed schema anchored on IFRS S2, with crosswalks to ESRS E1 and GRI 305 (Table 2), so that one extraction serves several regimes. Each field records its value, unit, scope and boundary qualifiers (for example, location- versus market-based Scope 2), reporting period, provenance and confidence. For physical risk, the schema records the hazard type, the disclosure level of Definition 2 and any named sites, facilities or locations, which are the raw material of the geo lane.

4.4 Distant supervision from XBRL-tagged filings

Annotated training data are scarce, especially outside English. Some regimes, however, publish the same information twice: as tagged XBRL facts and as a rendered document. Indian BRSR filings are one example. W2C aligns each tagged fact with its rendered occurrence (Algorithm 1) to create silver labels, following the logic of distant supervision (Mintz et al., 2009). Alignment combines numeric normalisation across scale factors and rounding, agreement of units and periods, and semantic similarity between the taxonomy label and the value's context: row and column headers for table cells, and the enclosing sentence for prose. Ambiguous alignments, in which a value occurs several times with similar scores, are discarded rather than guessed. A large teacher model, prompted with schema definitions and validated against the silver labels, then labels documents in languages without tagged filings, and a smaller multilingual

Algorithm 1 Distant supervision from XBRL-tagged filings

Input: Filing f with tagged facts $\{(c_j, v_j, u_j, p_j)\}$ (concept, value, unit, period); parsed rendering R_f with provenance; thresholds θ, μ
Output: Silver-labelled spans S_f

$S_f \leftarrow \emptyset$
for each fact (c, v, u, p) **do**
  $C \leftarrow \{s \in R_f : |\mathrm{num}(s) \cdot 10^k - v| \le \mathrm{tol}(s) \text{ for some } k \in \{0, \pm 3, \pm 6\}\}$
  for each candidate span $s \in C$ **do**
    $\mathrm{score}(s) \leftarrow \lambda_1 \,\mathrm{sim}\big(\mathrm{label}(c), \mathrm{ctx}(s)\big) + \lambda_2 \,\mathbf{1}[\mathrm{unit}(s) \simeq u] + \lambda_3 \,\mathbf{1}[\mathrm{period}(s) \simeq p]$
  $s^\star \leftarrow \arg\max_{s \in C} \mathrm{score}(s)$; $\Delta \leftarrow \mathrm{score}(s^\star) - \max_{s \in C \setminus \{s^\star\}} \mathrm{score}(s)$
  if $\mathrm{score}(s^\star) \ge \theta$ **and** $\Delta \ge \mu$ **then**
    $S_f \leftarrow S_f \cup \{(s^\star, c, v, u, p, \mathrm{score}(s^\star))\}$ ▷ ambiguous matches are discarded
return S_f

student model is distilled (Hinton et al., 2015) from teacher outputs that pass verification. Distillation lowers the cost and energy of processing global corpora.

4.5 Verification and calibration

Extracted values pass deterministic checks—unit normalisation, recomputation of totals (for example, Scope 1 plus Scope 2 against a reported total) and year-on-year plausibility bounds—as well as model-based checks. For the ordinal disclosure level, W2C estimates $P(D \mid \hat{D})$ on a held-out, double-annotated calibration sample and reports split-conformal prediction sets with guaranteed marginal coverage (Angelopoulos and Bates, 2023); numeric fields receive conformal intervals around their normalised values. The calibrated distribution, not the point prediction, enters the propagation step.

4.6 Entity resolution and geocoding with precision classes

Firms are resolved to Legal Entity Identifiers and parent–subsidiary relationships (Global Legal Entity Identifier Foundation, 2024), so that assets held by subsidiaries are attributed to the reporting group. Asset mentions extracted from reports are matched to open asset registries (Byers et al., 2018; Global Energy Monitor, 2024; Climate TRACE, 2024) and to OpenStreetMap features; unmatched mentions are geocoded. Every location carries a precision class $r \in \{\text{site, facility-name match, industrial zone, municipality, province, country}\}$ and an error scale σ_r estimated empirically against registries with known coordinates. Where only an administrative unit is known, location draws can be sampled uniformly within its polygon instead of from a Gaussian.

4.7 Hazard overlay

Locations are overlaid with hazard layers for each hazard, scenario (for example, SSP1-2.6, SSP2-4.5 and SSP5-8.5) and horizon (for example, 2030 and 2050). W2C keeps ensemble members separate rather than averaging them, so that disagreement between hazard models is propagated rather than hidden. Aggregation to H3 hexagonal cells (Uber Technologies, 2018) provides a common spatial index for storage, joins and maps.

4.8 Uncertainty propagation

Algorithm 2 draws, in each Monte Carlo iteration, a hazard-model member, a location for every asset and a disclosure level for every firm; it then recomputes exposures and peer ranks and records each firm's quadrant and DGI. The output is a probability for each quadrant and an interval for the DGI. Computation is embarrassingly parallel across draws. In our simulation, 80 draws for 1,247 firms and 11,107 assets take about ten seconds on a single machine, dominated by evaluating the synthetic hazard fields; with raster hazard layers, location draws reduce to array look-ups.

4.9 Delivery

The map aggregates firm- and asset-level results to H3 cells, offers scenario and horizon toggles, displays uncertainty explicitly, for example through opacity proportional to classification confidence, and links every asset to the report passage that mentions it. Published data comprise extracted values, provenance and uncertainty, following the FAIR principles (Wilkinson et al., 2016).

Algorithm 2 Monte Carlo propagation of extraction, geolocation and hazard-model uncertainty

Input: Extracted levels $\hat{D}_i$ and calibrated $P(D \mid \hat{D})$; locations $\hat{\mathbf{x}}_a$ with scales $\sigma_{r(a)}$; weights w_a; ensemble $\{Z^{(m)}\}_{m=1}^{M}$; thresholds τ_E, τ_D; number of draws K

Output: $\hat{P}(Q_i = q)$ for all firms i and quadrants q; draws $\{\mathrm{DGI}_i^{(k)}\}$

for $k = 1, \ldots, K$ **do**
- draw $m_k \sim \mathrm{Uniform}\{1, \ldots, M\}$
- draw $\mathbf{x}_a^{(k)} \sim \mathcal{N}\big(\hat{\mathbf{x}}_a, \sigma_{r(a)}^2 I_2\big)$ for every asset ▷ or uniformly within an administrative polygon
- $E_i^{(k)} \leftarrow \sum_{a \in \mathscr{A}_i} w_a \, g\big(Z^{(m_k)}(\mathbf{x}_a^{(k)})\big)$ for every firm
- draw $D_i^{(k)} \sim P(D \mid \hat{D}_i)$ for every firm
- compute $F_E^{(k)}$ and $F_D^{(k)}$ within the comparison set
- $Q_i^{(k)} \leftarrow$ quadrant from (3); $\quad \mathrm{DGI}_i^{(k)} \leftarrow F_E^{(k)}(E_i^{(k)}) - F_D^{(k)}(D_i^{(k)})$

return $\hat{P}(Q_i = q) = K^{-1} \sum_k \mathbf{1}[Q_i^{(k)} = q]$ and $\{\mathrm{DGI}_i^{(k)}\}_{k=1}^{K}$

5 Simulation study

We study the measurement properties of the DGI in a setting where the ground truth is known. Asset geography is real; hazards, disclosures and extraction errors are simulated. Table 5 in Appendix B lists all parameters.

5.1 Design

Assets and firms. We use version 1.3.0 of the Global Power Plant Database (Byers et al., 2018), which lists 34,936 plants with coordinates and capacity (Figure 4). We construct pseudo-firms by grouping plants by the database's owner field and retaining owners with at least three plants, which leaves 11,107 plants held by 1,247 owners in 76 countries. Weights w_a are capacity shares within each owner. Owner names are used only for grouping; no result in this paper refers to any real company.

Hazards. Hazard fields are unit-variance Gaussian random fields on the sphere with a squared-exponential kernel in chordal distance, sampled with 1,024 random Fourier features (Rahimi and Recht, 2007; Rasmussen and Williams, 2006) (Appendix B), for correlation lengths ℓ from 10 to 1,000 km. We set $g = \Phi$, the standard normal distribution function. Figure 2 shows one realisation at three correlation lengths over Thailand.

Disclosure and extraction. True disclosure levels follow a latent-awareness model, $L_i = \beta \, \Phi^{-1}\big(F_E(E_i)\big) + \sqrt{1-\beta^2}\, u_i$ with $u_i \sim \mathcal{N}(0,1)$, discretised into five levels with population shares $(0.15, 0.30, 0.25, 0.20, 0.10)$. With $\beta = 0.4$ the model produces a realistic mix of quadrants: on average 29% of firms are benign, 21% cheap talk, 16% silent risk and 34% aware. Extraction errors move a level to an adjacent level with probability ε. Geolocation errors are Gaussian with scale σ or, for country-level geocoding, replace each asset's location by the label point of its country (Natural Earth, 2024).

5.2 E1: Geolocation precision versus hazard scale

For each ℓ we draw 30 fields and compare exposures computed at true and degraded locations for seven error scales and for country-level geocoding. Figure 5 plots asset-level Spearman correlations against σ/ℓ. All 49 combinations of ℓ and σ collapse onto the curve predicted by Proposition 1, with a mean absolute deviation of 0.002. The largest deviation, 0.024, occurs at the longest correlation lengths, where assets are clustered relative to ℓ (the mean pairwise correlation $\bar{c}$ reaches 0.13 at $\ell = 1{,}000$ km); the finite-domain correction of Remark 1 reduces the mean deviation further to 0.0019.

Firm-level fidelity (Figure 6) is somewhat higher than asset-level fidelity at the same σ/ℓ—between 0.53 and 0.63 at $\sigma = \ell$, compared with about 0.48—because averaging over a firm's assets dampens independent location errors. Two results stand out. First, country-level geocoding destroys most of the signal: firm-level rank correlations are at most 0.11 for $\ell \leq 100$ km and only 0.29–0.54 for ℓ between 250 and 1,000 km. Second, city-level geolocation ($\sigma = 25$ km) is adequate for regional hazards, with rank correlations of at least 0.99 for $\ell \geq 250$ km, but not for local ones (0.57 at $\ell = 25$ km).

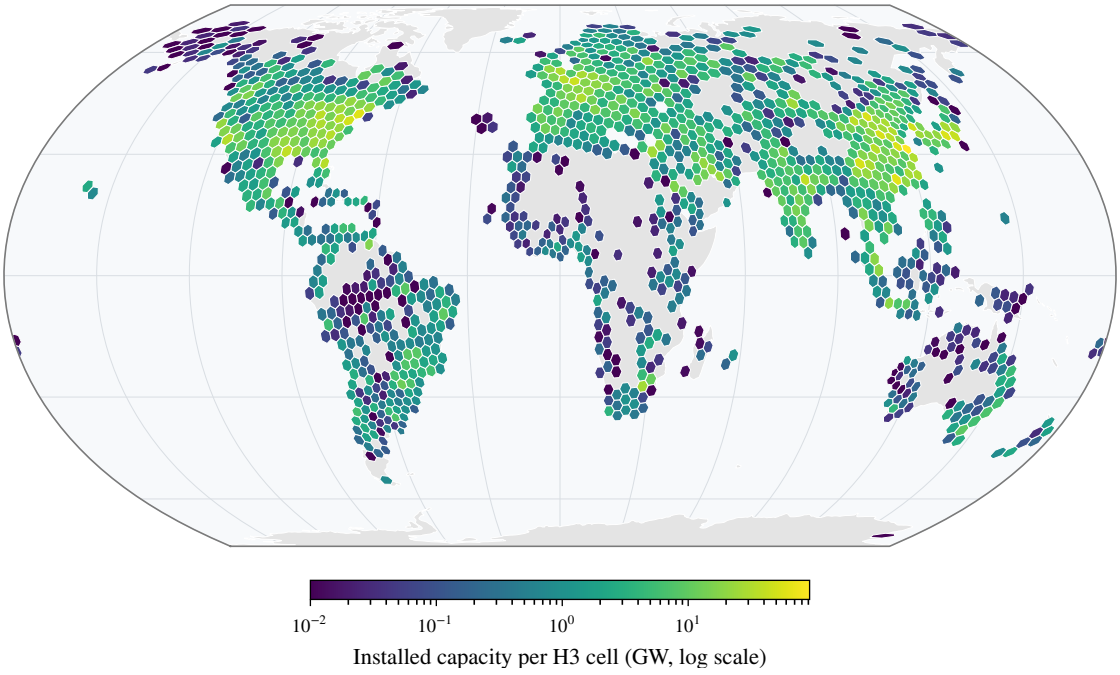


Figure 4. Asset geography used in the simulation: installed capacity of the 34,936 plants in the Global Power Plant Database v1.3.0 (Byers et al., 2018), aggregated to H3 resolution-2 cells (Uber Technologies, 2018) (mean cell area about 87,000 km^2) and shown in the Equal Earth projection (Šavrič et al., 2019). The simulation uses the 11,107 plants whose owners hold at least three plants.

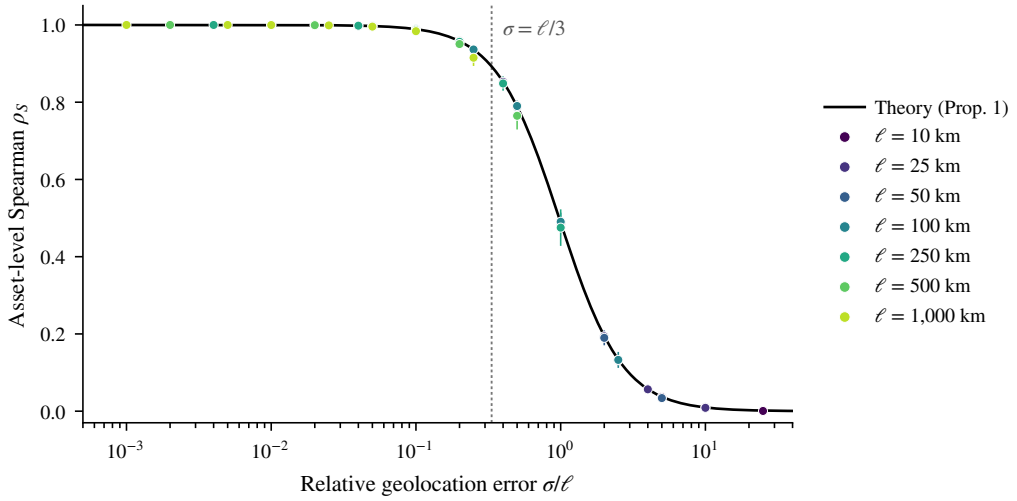


Figure 5. Asset-level fidelity of exposure under Gaussian geolocation error: Spearman correlation between hazard values at true and perturbed locations of 11,107 plants, plotted against σ/ℓ for seven correlation lengths and seven error scales (mean $\pm$ s.d. over 30 fields). The black curve is Equation (5); the dotted line marks the design rule $\sigma = \ell/3$.

5.3 E2: Extraction and geolocation errors in quadrant classification

Figure 7 reports four-class quadrant accuracy, measured against quadrants computed from true exposure and true disclosure, for a local ($\ell = 25$ km) and a regional ($\ell = 500$ km) hazard. Extraction error reduces accuracy almost linearly, by about 2.75 percentage points per 0.1 increase in ε, for both hazards. The slope has a simple explanation: an adjacent-level error changes the quadrant only when it crosses τ_D, which happens for levels 1 and 2 with probability $\varepsilon/2$ each, giving an expected loss of $\varepsilon(\pi_1 + \pi_2)/2 = 0.275\,\varepsilon$ for the level shares π used here.

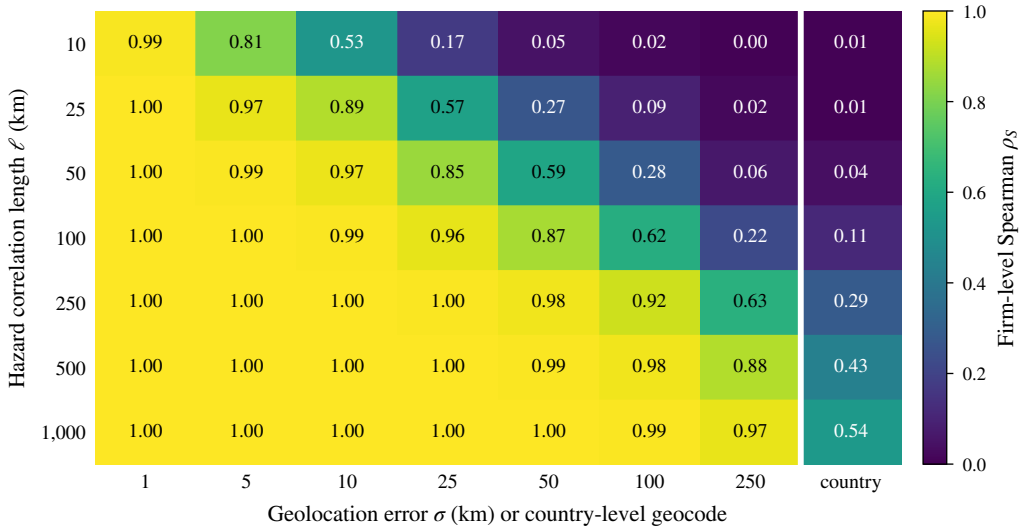


Figure 6. Firm-level fidelity: Spearman correlation between capacity-weighted firm exposures computed at true and degraded locations (1,247 pseudo-firms; mean over 30 fields). The rightmost column replaces every asset's location by its country's label point.

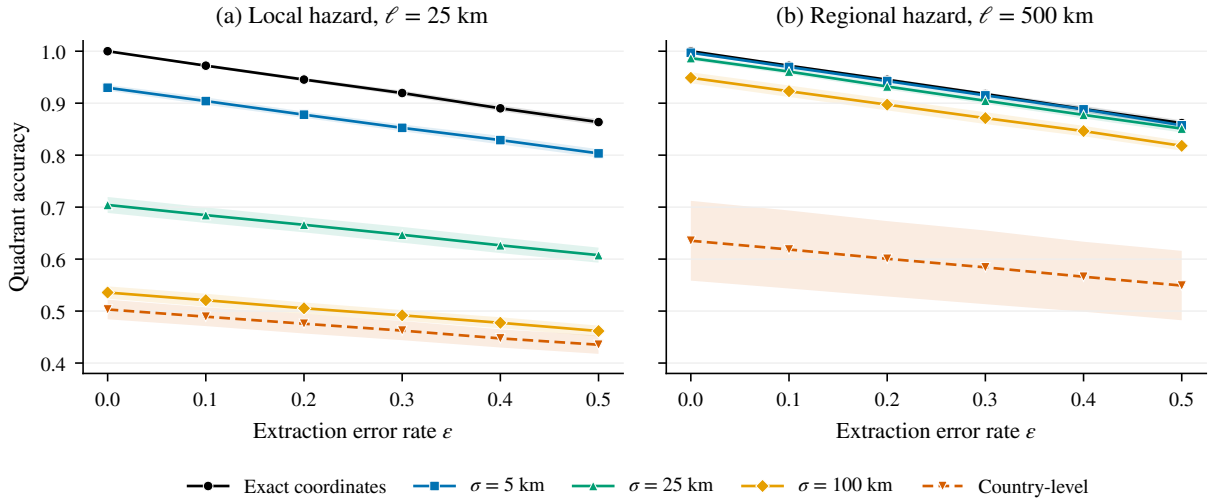


Figure 7. Four-class quadrant accuracy as a function of the extraction error rate ε for a local (a) and a regional (b) hazard, under exact, Gaussian-perturbed (σ = 5, 25 and 100 km) and country-level asset locations. Lines show means and bands ± 1 s.d. over 30 replications.

Geolocation error, in contrast, depends strongly on the hazard. For the local hazard, city-level error ($\sigma = 25$ km) alone reduces accuracy from 1.00 to 0.70, a loss more than five times larger than that from a 20% extraction error rate (accuracy 0.945), and reduces the F1 score of the silent-risk class to 0.62; country-level geocoding brings accuracy down to 0.50. For the regional hazard, the same city-level error costs just over one percentage point (0.987), whereas country-level geocoding still costs about 36 points (0.635). The two error sources combine approximately additively; for the local hazard with $\sigma = 25$ km and $\varepsilon = 0.2$, accuracy is 0.666. The practical implication is a hazard-specific allocation of engineering effort: for local hazards, investment in geolocation dominates; for regional hazards, extraction quality matters more.

5.4 E3: Uncertainty propagation

We next ask whether propagating uncertainty improves decisions. We fix a flood-like hazard ($\ell = 50$ km), city-level geolocation error ($\sigma = 25$ km), an extraction error rate of $\varepsilon = 0.25$, and hazard-model uncertainty represented by a

Table 3. Uncertainty propagation for identifying silent-risk firms (E3): means over 20 replications, standard deviations in parentheses. Brier scores and expected calibration error (ECE) evaluate $P(Q_i = \text{silent})$; AP is average precision. Precision, recall and the share of firms flagged refer to the point classification for the point estimate and to the rule $P(Q_i = \text{silent}) \geq 0.8$ for the Monte Carlo (MC) variants. Bold marks the best value in each probabilistic or precision column.

	Probabilistic quality			Flagging silent-risk firms		
Method	Brier ↓	ECE ↓	AP ↑	Precision ↑	Recall	Flagged
Point estimate	0.100 (.008)	–	–	0.676 (.029)	0.728 (.038)	17.4%
MC: extraction only	0.091 (.007)	0.075 (.009)	0.573 (.041)	0.675 (.030)	0.707 (.057)	16.9%
MC: location + hazard model	0.075 (.005)	0.047 (.007)	0.712 (.032)	0.803 (.038)	0.429 (.034)	8.6%
MC: full propagation	**0.072** (.004)	**0.033** (.006)	**0.732** (.026)	**0.842** (.029)	0.299 (.057)	5.7%

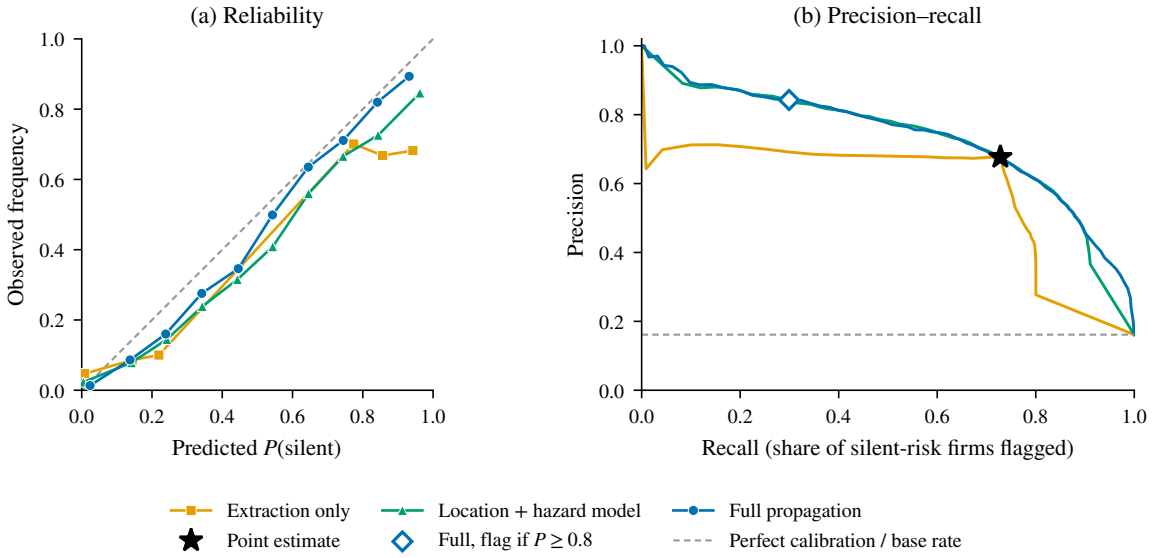


Figure 8. Uncertainty propagation for identifying silent-risk firms ($\ell = 50$ km, $\sigma = 25$ km, $\varepsilon = 0.25$, five-member hazard ensemble; 20 replications pooled). (a) Reliability of predicted probabilities, for bins containing at least 30 firms. (b) Precision–recall curves; the star marks the point estimate and the diamond the rule that flags a firm when $P(\text{silent}) \geq 0.8$ under full propagation. The dashed line in (b) is the base rate of 16.1%.

five-member ensemble whose members correlate with the true field at $\sqrt{1-\kappa^2} \approx 0.92$. The predictive distribution $P(D \mid \hat{D})$ is estimated on a simulated calibration sample of 500 annotated reports. We compare a point estimate, which uses the ensemble mean at the estimated locations and the extracted level, with three Monte Carlo variants that propagate extraction uncertainty only, location and hazard-model uncertainty only, or all three jointly. Silent-risk firms make up 16.1% of firms on average.

Table 3 and Figure 8 summarise the results. Full propagation gives the best probabilistic predictions: the Brier score falls from 0.100 for the point estimate to 0.072, and expected calibration error falls from 0.075 with extraction-only propagation to 0.033. Ignoring location and hazard-model uncertainty makes probabilities overconfident: firms to which extraction-only propagation assigns $P(\text{silent}) \geq 0.8$ are silent in only about 68% of cases.

Propagation does not create information. At the point estimate's recall of 0.73, the propagated methods and the point estimate achieve essentially the same precision (0.67–0.68). What propagation adds is the ability to use the information selectively: the precision–recall curve of full propagation dominates that of extraction-only propagation, and a rule that flags a firm only when $P(\text{silent}) \geq 0.8$ attains a precision of 0.84, against 0.68 for the point estimate, while flagging 5.7% of firms and recovering 30% of silent-risk firms. Where a false finding of silence is costly, as in supervisory triage or public reporting, this abstention-based operating point is preferable to a point classification.

Table 4. Evaluation protocol for the components of W2C.

Component	Reference data	Metrics
Disclosure level D	Double-annotated gold set, stratified by jurisdiction, language and sector	Macro-F1; quadratic-weighted κ; Krippendorff's α between annotators; ECE; conformal coverage and set size
Numeric metrics	XBRL facts (e.g., BRSR); CDP responses where licensed	Exact match; relative error $\leq 1\%$; accuracy of units, scope and boundary qualifiers
Provenance	Gold page and bounding box	Page accuracy; bounding-box intersection over union
Asset geolocation	Registry coordinates (GPPD, GEM, Climate TRACE)	Median error (km); share within 10 and 25 km; calibration of σ_r
Hazard exposure	Several hazard products and ensemble members	Cross-product rank agreement
End to end	Expert review of a stratified sample of firms	Quadrant agreement; precision of silent-risk flags at fixed coverage

6 Evaluation protocol for the empirical programme

The simulation establishes how the DGI behaves when the error processes are known; the empirical programme must measure those processes. Table 4 lists the evaluation of each component. Three design choices deserve emphasis. First, the gold standard is stratified by jurisdiction, language and sector, so that performance in under-represented markets is measured rather than assumed, and inter-annotator agreement on disclosure levels is reported with Krippendorff's α (Krippendorff, 2004). Second, geolocation error scales σ_r are estimated against registries with known coordinates; Proposition 1 shows that these scales, relative to each hazard's correlation length, decide whether exposure estimates are informative at all. Third, exposure is computed from more than one hazard product, so that the disagreement documented by Hain et al. (2022) becomes part of the reported uncertainty rather than a hidden modelling choice.

7 Research agenda and identification strategies

W2C is designed to answer three empirical questions.

RQ1. *Measurement.* How accurately can taxonomy-aligned disclosures and asset locations be extracted across languages and reporting regimes, and how much does distant supervision from XBRL improve performance in languages without labelled data?

RQ2. *Description.* How large is the gap between disclosed and modelled physical risk, and how does it vary across hazards, sectors and countries? Quadrant probabilities aggregated on H3 cells provide the descriptive basis.

RQ3. *Explanation and consequences.* Do mandatory disclosure regimes close the gap, and do markets price undisclosed exposure?

For RQ3, the staggered adoption of ISSB-aligned requirements across jurisdictions provides variation in treatment timing that can be analysed with heterogeneity-robust difference-in-differences estimators (Callaway and Sant'Anna, 2021), with the firm-level DGI as outcome and firms in jurisdictions not yet treated as comparisons. Threats include anticipation, selection into early adoption and concurrent reforms; event-time plots and placebo outcomes address them in part. A complementary design exploits realised hazard events: when the footprint of a flood or cyclone intersects a firm's assets, an event study (MacKinlay, 1997) can test whether abnormal returns differ between firms that had disclosed the exposure and firms classified as silent. Because the DGI is estimated with error, regressions should use its full posterior—through multiple imputation over Monte Carlo draws or errors-in-variables corrections (Carroll et al., 2006)—rather than point estimates.

Thailand offers a natural deep-dive case. Listed Thai companies report through the Form 56-1 One Report in Thai and English, which tests multilingual extraction; power and industrial assets cluster in the Chao Phraya basin and along the Eastern Seaboard (Figure 2); and the 2011 floods, which disrupted manufacturing and global supply chains, provide a historical shock against which exposure estimates can be validated. Because the framework is global by design, lessons from the deep dive transfer to other markets with similar data constraints.

8 Discussion

Limitations of the simulation. The simulation abstracts from features of real data. Hazard fields are stationary and Gaussian, whereas real hazards are non-stationary and heavy-tailed; location errors are Gaussian, whereas geocoding

errors are structured by administrative boundaries; and disclosure behaviour follows a stylised latent-awareness model. The asset sample comes from a single sector, and its owner field over-represents jurisdictions with open power-sector data, notably the United States. These choices make the ground truth known and the theory testable; they do not substitute for the empirical evaluation of Section 6.

Exposure is not risk. The DGI compares disclosure with modelled hazard exposure, not with vulnerability, adaptation or financial materiality; a firm whose assets in a flood plain are well protected may rationally disclose little. Asset weights should therefore reflect economic significance where it can be measured, and quadrant labels should be read as prompts for scrutiny rather than findings of non-compliance. Asset coverage is also incomplete: registries cover energy and heavy industry far better than services or agriculture, and firms' own location disclosures are uneven, which is itself an aspect of the gap that W2C measures.

Ethics and responsible use. Labelling a firm as silent risk has reputational consequences. W2C therefore reports probabilities rather than labels, abstains below a confidence threshold, exposes provenance so that any classification can be checked against its sources, and should give firms a channel to correct errors. Language models used for extraction can produce plausible but wrong values; verification, provenance and calibration are the safeguards, and no value enters the published dataset without a source location. Source reports are publicly available but protected by copyright, so W2C publishes extracted facts, hashes and links rather than documents; some hazard and biodiversity datasets restrict commercial use, which constrains the redistribution of derived products.

Environmental footprint. Processing tens of thousands of reports with large models consumes energy (Strubell et al., 2019). W2C reserves large models for teacher labelling and validation and uses distilled models for bulk processing, and we recommend reporting the computational footprint of the pipeline (Lacoste et al., 2019)—a modest but consistent requirement for research that evaluates other organisations' climate disclosures.

9 Conclusion

Climate disclosure is becoming abundant, standardised and consequential, but its connection to the physical world remains largely unmeasured. Words to Coordinates links what firms disclose to what their assets face, carrying provenance and uncertainty from source documents to a peer-relative Disclosure Gap Index and a global map. Our analysis yields a simple, testable design rule—geolocation error should be at most one third of a hazard's correlation length—and shows, on real asset geography, that country-level location information is inadequate for physical risk assessment, that geolocation rather than extraction is the binding constraint for local hazards, and that propagating uncertainty turns noisy measurements into calibrated, decision-ready probabilities. The next step is the empirical programme set out in Sections 6 and 7: building the multilingual benchmark, estimating real error scales and measuring the disclosure gap across the world's markets. Because location is also the first step of nature-related assessment under the TNFD's approach (Taskforce on Nature-related Financial Disclosures, 2023), the same architecture extends naturally from climate to nature.

Data and code availability

The simulation uses the Global Power Plant Database v1.3.0 (CC BY 4.0) (Byers et al., 2018) and Natural Earth boundaries (public domain) (Natural Earth, 2024). All other inputs are simulated with fixed random seeds. Code to reproduce every result and figure is available from the author on request.

Use of AI tools

Generative AI tools were used to assist with drafting text, writing code and preparing figures. The author reviewed the manuscript and takes full responsibility for its content.

References

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Christoph Auer et al. Docling technical report. *arXiv preprint arXiv:2408.09869*, 2024.

Florian Berg, Julian F. Kölbel, and Roberto Rigobon. Aggregate confusion: The divergence of ESG ratings. *Review of Finance*, 26(6):1315–1344, 2022.

Julia Anna Bingler, Mathias Kraus, Markus Leippold, and Nicolas Webersinke. Cheap talk and cherry-picking: What ClimateBert has to say on corporate climate risk disclosures. *Finance Research Letters*, 47:102776, 2022.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Logan Byers, Johannes Friedrich, Roman Hennig, Aaron Kressig, Xinyue Li, Colin McCormick, and Laura Malaguzzi Valeri. A global database of power plants. Technical note, World Resources Institute, 2018. Global Power Plant Database, version 1.3.0 (2021).

Ben Caldecott, Matthew McCarten, Christophe Christiaen, and Conor Hickey. Spatial finance: Practical and theoretical contributions to financial analysis. *Journal of Sustainable Finance & Investment*, 2022.

Brantly Callaway and Pedro H. C. Sant'Anna. Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230, 2021.

Raymond J. Carroll, David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman & Hall/CRC, second edition, 2006.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. BGE M3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 2024.

Climate TRACE. Climate TRACE emissions inventory. https://climatetrace.org, 2024.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. CLIMATE-FEVER: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*, 2020.

European Commission. Commission Delegated Regulation (EU) 2023/2772 supplementing Directive 2013/34/EU as regards sustainability reporting standards. Official Journal of the European Union, 2023.

Veronika Eyring, Sandrine Bony, Gerald A. Meehl, Catherine A. Senior, Bjorn Stevens, Ronald J. Stouffer, and Karl E. Taylor. Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958, 2016.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. ColPali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*, 2024.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Global Energy Monitor. Global Energy Monitor trackers. https://globalenergymonitor.org, 2024.

Global Legal Entity Identifier Foundation. LEI reference and relationship data. https://www.gleif.org, 2024.

Global Reporting Initiative. GRI 305: Emissions 2016. Global Sustainability Standards Board, 2016.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330, 2017.

Linda Isabella Hain, Julian F. Kölbel, and Markus Leippold. Let's get physical: Comparing metrics of physical climate risk. *Finance Research Letters*, 46:102406, 2022.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

International Sustainability Standards Board. IFRS S1 General Requirements for Disclosure of Sustainability-related Financial Information. IFRS Foundation, 2023a.

International Sustainability Standards Board. IFRS S2 Climate-related Disclosures. IFRS Foundation, 2023b.

Kenneth R. Knapp, Michael C. Kruk, David H. Levinson, Howard J. Diamond, and Charles J. Neumann. The International Best Track Archive for Climate Stewardship (IBTrACS): Unifying tropical cyclone data. *Bulletin of the American Meteorological Society*, 91(3):363–376, 2010.

Julian F. Kölbel, Markus Leippold, Jordy Rillaerts, and Qian Wang. Ask BERT: How regulatory disclosure of transition and physical climate risks affects the CDS term structure. *Journal of Financial Econometrics*, 22(1), 2024.

Sakis Kotsantonis and George Serafeim. Four things no one will tell you about ESG data. *Journal of Applied Corporate Finance*, 31(2):50–58, 2019.

Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology*. Sage, second edition, 2004.

Philipp Krueger, Zacharias Sautner, and Laura T. Starks. The importance of climate risks for institutional investors. *The Review of Financial Studies*, 33(3):1067–1111, 2020.

Samantha Kuzma et al. Aqueduct 4.0: Updated decision-relevant global water risk indicators. Technical note, World Resources Institute, 2023.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019.

Patrick Lewis et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020.

Alexandra Luccioni, Emily Baylor, and Nicolas Duchene. Analyzing sustainability reports using natural language processing. *arXiv preprint arXiv:2011.08073*, 2020.

A. Craig MacKinlay. Event studies in economics and finance. *Journal of Economic Literature*, 35(1):13–39, 1997.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, 2009.

Natural Earth. Natural Earth: Free vector and raster map data. https://www.naturalearthdata.com, 2024.

Network for Greening the Financial System. NGFS climate scenarios for central banks and supervisors. NGFS, 2023.

Jingwei Ni et al. CHATREPORT: Democratizing sustainability disclosure analysis through LLM-based tools. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2023.

Brian C. O'Neill et al. The Scenario Model Intercomparison Project (ScenarioMIP) for CMIP6. *Geoscientific Model Development*, 9:3461–3482, 2016.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

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

Zacharias Sautner, Laurence van Lent, Grigory Vilkov, and Ruishen Zhang. Firm-level climate change exposure. *The Journal of Finance*, 78(3):1449–1498, 2023.

Bojan Šavrič, Tom Patterson, and Bernhard Jenny. The Equal Earth map projection. *International Journal of Geographical Information Science*, 33(3):454–465, 2019.

Tobias Schimanski, Andrin Reding, Nico Reding, Julia Bingler, Mathias Kraus, and Markus Leippold. Bridging the gap in ESG measurement: Using NLP to quantify environmental, social, and governance communication. *Finance Research Letters*, 61:104979, 2024a.

Tobias Schimanski et al. ClimRetrieve: A benchmarking dataset for information retrieval from corporate climate disclosures. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024b.

Dominik Stammbach, Nicolas Webersinke, Julia Anna Bingler, Mathias Kraus, and Markus Leippold. Environmental claim detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2023.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, 2019.

Task Force on Climate-related Financial Disclosures. Final report: Recommendations of the Task Force on Climate-related Financial Disclosures. Financial Stability Board, 2017.

Taskforce on Nature-related Financial Disclosures. Recommendations of the Taskforce on Nature-related Financial Disclosures. TNFD, 2023.

Bridget Thrasher, Weile Wang, Andrew Michaelis, Forrest Melton, Tsengdar Lee, and Ramakrishna Nemani. NASA global daily downscaled projections, CMIP6. *Scientific Data*, 9:262, 2022.

Uber Technologies. H3: A hexagonal hierarchical geospatial indexing system. https://h3geo.org, 2018.

Francesco S. Varini, Jordan Boyd-Graber, Massimiliano Ciaramita, and Markus Leippold. ClimaText: A dataset for climate change topic detection. *arXiv preprint arXiv:2012.00483*, 2020.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Philip J. Ward et al. Aqueduct Floods methodology. Technical note, World Resources Institute, 2020.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. ClimateBert: A pretrained language model for climate-related text. *arXiv preprint arXiv:2110.12010*, 2021.

Mark D. Wilkinson et al. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018, 2016.

A Proof of Proposition 1 and Remark 1

Pearson correlation. Conditional on $\boldsymbol{\delta}$, stationarity and isotropy imply that $\big(Z(\mathbf{x}), Z(\mathbf{x}+\boldsymbol{\delta})\big)$ is bivariate normal with unit variances and correlation $\rho(\|\boldsymbol{\delta}\|)$. Because $\boldsymbol{\delta}$ is independent of the field, $\mathrm{Cov}\big(Z(\mathbf{x}), Z(\hat{\mathbf{x}})\big) = \mathbb{E}_{\boldsymbol{\delta}}\big[\rho(\|\boldsymbol{\delta}\|)\big]$, and both marginal variances equal one. With $\boldsymbol{\delta} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_2)$, $\|\boldsymbol{\delta}\|^2/\sigma^2 \sim \chi^2_2$, so that $T := \|\boldsymbol{\delta}\|^2/(2\sigma^2)$ is standard exponential and $\rho(\|\boldsymbol{\delta}\|) = \exp\{-(\sigma^2/\ell^2)\,T\}$. Hence

$$\mathbb{E}\big[\rho(\|\boldsymbol{\delta}\|)\big] = \int_0^\infty e^{-(\sigma^2/\ell^2)t}\, e^{-t}\, dt = \frac{1}{1+\sigma^2/\ell^2} = \frac{\ell^2}{\ell^2+\sigma^2},$$

which proves (4).

Spearman correlation. Spearman's correlation is invariant under strictly increasing transformations, so we may take g to be the identity. For a pair (X,Y) with continuous marginals, the population Spearman correlation is $\rho_S = 3\big[P\{(X_1-X_2)(Y_1-Y_3)>0\} - P\{(X_1-X_2)(Y_1-Y_3)<0\}\big]$, where (X_j,Y_j), $j=1,2,3$, are independent copies of (X,Y). Here $X = Z(\mathbf{x}_a)$ and $Y = Z(\hat{\mathbf{x}}_a)$, and independent copies correspond to independent assets. The marginals of X and Y are standard normal whatever the value of $\boldsymbol{\delta}$, so X_2 and Y_3 are independent standard normal variables. Conditional on the displacement $\boldsymbol{\delta}_1$ of the first copy, (X_1,Y_1) is bivariate normal with correlation $\rho(\|\boldsymbol{\delta}_1\|)$, and the bracket equals the bivariate-normal value $(6/\pi)\arcsin\{\rho(\|\boldsymbol{\delta}_1\|)/2\}$. The expression is linear in the distribution of (X_1,Y_1), so averaging over $\boldsymbol{\delta}_1$, with the substitution above, gives (5). Both (4) and (5) depend on (σ,ℓ) only through σ/ℓ. □

Finite domain. For n assets observed in a single realisation of the field, the expected within-sample covariance between $Z(\mathbf{x}_a)$ and $Z(\hat{\mathbf{x}}_a)$ is approximately $\rho - \bar{c}$, and the expected within-sample variance of $Z(\mathbf{x}_a)$ is approximately $1-\bar{c}$, where $\bar{c} = n^{-2}\sum_{a,b}\rho(d_{ab})$ is the mean pairwise correlation and the approximation neglects terms of order $1/n$ and differences between $\rho(d_{ab})$ and $\rho(\|\mathbf{x}_a - \hat{\mathbf{x}}_b\|)$. The within-sample correlation is therefore approximately $(\rho-\bar{c})/(1-\bar{c})$. For the plant sample, $\bar{c}$ ranges from 0.0004 at $\ell = 10$ km to 0.13 at $\ell = 1{,}000$ km.

B Simulation details

Random fields. Hazard fields are generated as

$$Z(\mathbf{x}) = \sqrt{2/J}\sum_{j=1}^{J}\cos\big(\boldsymbol{\omega}_j^\top \mathbf{p}(\mathbf{x}) + b_j\big), \qquad \boldsymbol{\omega}_j \sim \mathcal{N}(\mathbf{0}, \ell^{-2} I_3), \quad b_j \sim \mathrm{Uniform}[0, 2\pi),$$

where $\mathbf{p}(\mathbf{x}) \in \mathbb{R}^3$ is the Earth-centred Cartesian position, in kilometres, of $\mathbf{x}$ on a sphere of radius 6,371 km, and $J = 1{,}024$. As $J \to \infty$, Z converges to a Gaussian process with covariance $\exp\{-\|\mathbf{p}(\mathbf{x}) - \mathbf{p}(\mathbf{x}')\|^2/(2\ell^2)\}$ (Rahimi and Recht, 2007). Restricted to the sphere, this is a valid covariance in chordal distance, which is approximately equal to great-circle distance at the scales considered.

Location errors. Gaussian perturbations are applied in the local tangent plane, converting kilometres to degrees of latitude and longitude at each asset's latitude. Country-level geocoding uses the Natural Earth label point of the asset's country; for the single country code in the database without a match, the mean location of that country's plants is used.

Ensemble members. Hazard-model members are $Z^{(m)} = \sqrt{1-\kappa^2}\,Z + \kappa\,\eta_m$, where the η_m are independent fields with the same correlation length as Z. The point estimate uses the ensemble mean.

Reproducibility. Each experimental unit (a correlation length in E1 and E2, a replication in E3) uses its own random-number stream derived from a fixed master seed, so results do not depend on the order or partitioning of computation. Evaluation metrics follow standard definitions: expected calibration error uses ten equal-width bins, and average precision is computed with scikit-learn.

Table 5. Parameters of the simulation study. E1–E3 denote the three experiments.

Component	Setting
Assets	GPPD v1.3.0; 11,107 plants of 1,247 owners with ≥ 3 plants, 76 countries; w_a = capacity share
Hazard fields	Gaussian random fields, squared-exponential kernel, 1,024 random Fourier features; $g = \Phi$
Correlation length ℓ (km)	E1: 10, 25, 50, 100, 250, 500, 1,000; E2: 25, 500; E3: 50
Geolocation error σ (km)	E1: 1, 5, 10, 25, 50, 100, 250, country level; E2: 0, 5, 25, 100, country level; E3: 25
Disclosure model	$\beta = 0.4$; level shares $(0.15, 0.30, 0.25, 0.20, 0.10)$
Extraction error ε	E2: 0, 0.1, …, 0.5; E3: 0.25
Quadrant thresholds	$\tau_E = 0.5$, $\tau_D = 2$
Hazard-model ensemble	E3: $M = 5$ members, $Z^{(m)} = \sqrt{1-\kappa^2}\, Z + \kappa\, \eta_m$ with $\kappa = 0.4$
Calibration of $P(D \mid \hat{D})$	E3: 500 simulated annotated reports, add-one smoothing
Monte Carlo	E3: $K = 80$ draws
Replications	E1: 30 fields per ℓ; E2: 30 per ℓ, three extraction draws each; E3: 20