Vol. 3, No. 3, pp.: 60-75, 2025

# OceanPlasticDB: A Marine Plastic Observation Database for Environmental AI and Policy Evaluation

**Aoife Brennan[1], Lars Eriksson[2], Fiona MacLachlan[3, *]**

[1] Department of Marine Science and Environmental Studies, University of Galway, Galway H91 TK33, Ireland

[2] Department of Physical Geography and Ecosystem Science, Lund University, 223 62 Lund, Sweden

[3] School of Geography and Sustainable Development, University of Dundee, Dundee DD1 4HN, UK

** f.z.maclachlan@dundee.ac.uk*

| Article Information | |
|---|---|
| Received | 18 July 2025 |
| Accepted | 29 August 2025 |
| DOI | https://doi.org/10.63646/datamind.2025.030305 |

## Abstract

Marine plastic pollution is recognised as a global environmental crisis, yet evidence-based policy design and AI-driven monitoring systems are constrained by the severe fragmentation and inconsistency of existing observational datasets. Plastic abundance records derived from beach surveys, drone overpasses, vessel trawls, satellite remote sensing, and port activity logs are dispersed across incompatible registries, differ in spatial resolution, depth coverage, temporal cadence, and polymer-classification schemes, and lack the linked governance metadata required to evaluate the effectiveness of regulatory interventions. This paper introduces OceanPlasticDB, an open, schema-documented, multi-source marine plastic observation database integrating 593,500 georeferenced records from six primary observational programmes spanning 1990 to 2024 across the global ocean. The database resolves three systematic deficiencies of existing data products: (i) a hierarchical label-harmonisation pipeline standardises plastic-type annotations across 12 polymer classes using ISO 472 nomenclature and YOLO-v8 confidence-weighted relabelling; (ii) ocean-state covariates (HYCOM sea-surface currents, ERA5 winds, AIS port activity indices) are collocated to each observation to support drift modelling and source attribution; and (iii) a structured policy register encoding 214 national and supranational plastic governance events (bans, levies, extended producer responsibility schemes) enables pre/post intervention analysis. The storage architecture comprises PostGIS for georeferenced observations, TimescaleDB for time-series ocean state data, Apache Parquet for drone and satellite image tiles, and Neo4j for river-to-coast plastic source-to-sink pathway graphs. Experimental validation demonstrates that a YOLO-v8 hotspot detector trained on OceanPlasticDB achieves F1 = 0.891, outperforming single-

source baselines by up to 25 percentage points. A Lagrangian drift model fine-tuned with database-collocated current features reduces 72-hour trajectory RMSE from 15.7 km to 7.3 km. A difference-in-differences policy analysis of 38 port single-use ban events documents a statistically significant 28% plastic density reduction ($p = 0.003$) in treated coastal zones within 24 months. The full database, pipeline code, and benchmark harness are released under CC BY 4.0 with a persistent DOI.

**Keywords:** ***Marine plastic pollution; environmental database; remote sensing; drift modelling; policy evaluation; postGIS; TimescaleDB; neo4j; YOLO; reproducible AI***

## 1. Introduction

The global ocean is estimated to contain between 75 and 199 million tonnes of plastic waste, with annual inputs of 8 to 12 million tonnes discharged primarily through riverine pathways, coastal communities, and maritime activities (Jambeck et al., 2015; van Sebille et al., 2020). This pollution constitutes a pervasive environmental stressor that affects marine biodiversity through ingestion, entanglement, and the leaching of persistent organic pollutants (Laist, 1997; Cole et al., 2011; Rochman et al., 2013). Despite growing scientific and policy urgency—exemplified by the United Nations Environment Assembly's 2022 resolution to negotiate a legally binding global plastics treaty (UNEP, 2022)—the evidence base available to support AI-driven monitoring systems and rigorous policy evaluation remains chronically fragmented. Observations from beach surveys, trawl nets, drone overflights, satellite imagery, and port logbooks are distributed across dozens of separate registries maintained by national agencies, research consortia, and non-governmental organisations. These registries differ in spatial resolution, temporal coverage, polymer classification schemes, depth referencing conventions, and the availability of linked contextual metadata such as ocean currents, wind forcing, and governance events (Maximenko et al., 2019; Lebreton et al., 2018).

This fragmentation creates four concrete barriers to scientific progress. First, AI models for marine plastic hotspot detection cannot be trained on diverse, globally representative data because no unified georeferenced corpus exists; published detection benchmarks are therefore limited to small, site-specific datasets that do not generalise across ocean basins or observation methods (Politikos et al., 2021). Second, Lagrangian drift simulations—the primary tool for attributing floating plastic to source rivers and coastlines—require collocated ocean current and wind observations that are not linked to plastic abundance records in any existing database, forcing researchers to manually merge data products of different provenance and resolution (Lebreton et al., 2012; van Sebille et al., 2020; Maximenko et al., 2019). Third, the causal evaluation of plastic governance interventions (single-use bans, producer levies, port reception facility requirements) requires pre/post abundance records at treated and control locations that can only be assembled with considerable manual effort from siloed observation archives. Fourth, the absence of standardised data access interfaces prevents automated, reproducible benchmarking of environmental AI models across research groups (Lu, 2019; Zhang & Lu, 2021).

This paper addresses these barriers by introducing OceanPlasticDB, a curated, schema-documented, global marine plastic observation database that unifies 593,500 georeferenced records from six primary data sources. Our contributions are: (i) a 20-field canonical schema covering observation geometry, plastic type and mass, size class, ocean state covariates, port activity, and governance event

tags; (ii) a four-tier storage architecture integrating PostGIS, TimescaleDB, Apache Parquet, and Neo4j; (iii) a reproducible construction pipeline implementing label harmonisation, covariate collocation, drift pre-computation, and policy encoding; and (iv) experimental validation through three downstream tasks—hotspot detection, drift trajectory prediction, and policy difference-in-differences analysis—demonstrating the utility of the integrated database over single-source alternatives. The paper is organised as follows: Section 2 defines the database gap and use cases; Section 3 reviews related work; Section 4 presents data sources and schema; Section 5 describes the construction pipeline; Section 6 reports experiments; Section 7 addresses reproducibility; Section 8 states limitations; Section 9 concludes.

## 2. Database Gap and Use Cases

The community of researchers working on marine plastic monitoring, drift modelling, and governance evaluation currently relies on a patchwork of incompatible data products. OSPAR's beach survey archive (OSPAR, 2024) provides high-quality standardised litter counts from the North-East Atlantic but covers only a single ocean region and excludes microplastics smaller than 2.5 cm. LITTERBASE (Kölmel & Eckhardt, 2018) aggregates peer-reviewed global observations but lacks machine-readable field definitions and polymer confidence scores. The IMO's ShipPlast initiative (IMO, 2020) collects port-based plastic inventory data but does not link records to ambient ocean concentration observations. Satellite remote sensing products from Sentinel-2 and MODIS provide global spatial coverage but require substantial pre-processing to extract meaningful plastic abundance estimates, and derived products from different research groups are not harmonised to a common abundance metric (Biermann et al., 2020; Kikaki et al., 2022). The UNEP Global Plastics Outlook (UNEP, 2021) provides policy inventory data but does not link governance events to georeferenced abundance observations, making causal impact evaluation impossible without manual data assembly.

OceanPlasticDB fills this gap through four design choices that directly address the identified deficiencies. First, it merges six complementary data sources spanning 1990 to 2024, providing temporal depth sufficient for pre/post governance impact analysis of major regulatory milestones including the EU Single-Use Plastics Directive (2019), national plastic bag levies, and the adoption of IMO port reception facility guidelines. Second, it collocates ERA5 wind reanalysis and HYCOM ocean current reanalysis data to each observation record, enabling drift attribution without manual data fusion. Third, it links 214 governance events from the UNEP and EU policy registries to affected coastal zones through spatial join operations, enabling automated identification of treated and control observation sites for difference-in-differences analysis. Fourth, its PostGIS indexing and OGC-compliant REST interface enable spatial queries that are not possible with flat-file archives (Kölmel & Eckhardt, 2018; OSPAR, 2024; IMO, 2020).

The four primary use cases served by OceanPlasticDB are: (1) AI-based hotspot detection and marine litter density mapping, using the labelled drone and satellite imagery tiles to train and benchmark computer vision models; (2) Lagrangian drift trajectory prediction, using the covariate-enriched flow records to initialise and validate particle tracking simulations; (3) governance impact evaluation, using the linked policy register and pre/post abundance records to quantify the effectiveness of regulatory interventions through quasi-experimental methods; and (4) source attribution and river-to-ocean

pathway analysis, using the Neo4j source-to-sink graph to trace likely riverine origins of observed coastal and offshore plastic concentrations (Jambeck et al., 2015; Lebreton et al., 2012; Rochman et al., 2013).

## 3. Related Work

### *3.1 Marine Plastic Monitoring and Observational Datasets*

The scientific study of ocean plastic pollution was catalysed by reports of the North Pacific Garbage Patch (Moore et al., 2001) and the seminal global mass balance analysis of Jambeck et al. (2015), which estimated per-capita mismanaged plastic waste for 192 coastal nations. Subsequent surface drift modelling by Lebreton et al. (2012) and van Sebille et al. (2020) established the Lagrangian particle tracking framework as the primary tool for simulating plastic transport from land-based sources to ocean accumulation zones. Maximenko et al. (2019) reviewed the principal data types available for validating drift simulations—including surface drifter trajectories, trawl surveys, and beach cast studies—and documented the systematic gaps in spatial and temporal coverage that limit model calibration. The LITTERBASE synthesis (Kölmel & Eckhardt, 2018) was the first comprehensive effort to aggregate peer-reviewed observational data into a single searchable interface, but it lacks a formal schema specification, confidence scores, and covariate linkage. The EU Copernicus Marine Environment Monitoring Service has begun publishing satellite-derived floating debris products but these remain at early technology readiness levels and are not linked to in situ validation data (Biermann et al., 2020).

### *3.2 Computer Vision for Plastic Detection*

The application of deep convolutional neural networks and object detection architectures to marine litter identification in drone and satellite imagery has accelerated since the availability of labelled training datasets. Politikos et al. (2021) demonstrated that U-Net semantic segmentation achieves 0.78 F1 for beach litter mapping from UAV imagery, while Wolf et al. (2020) reported 0.83 F1 using Mask R-CNN on RGB drone captures of Mediterranean beach transects. Kikaki et al. (2022) applied MARIDA—a Sentinel-2 marine debris dataset—to train pixel-level classifiers for floating marine litter, achieving 0.71 overall accuracy on a 10-class scheme. These studies all suffer from the same limitation: they were trained on local, single-site datasets that do not generalise to other ocean regions or observation platforms. OceanPlasticDB's 62,700 drone UAV records and 187,300 processed satellite tiles, spanning multiple ocean basins and acquired with standardised observation protocols, directly address this generalisation bottleneck (Biermann et al., 2020; Lu, 2022; Lu & Xu, 2019).

### *3.3 Policy Evaluation Methods for Environmental Interventions*

Causal evaluation of environmental policies using observational data has advanced substantially through the adoption of quasi-experimental methods from econometrics, including difference-in-differences (DiD) estimators, regression discontinuity designs, and synthetic control methods (Angrist & Pischke, 2009). DiD is particularly well suited to evaluating plastic governance interventions because treated (regulated) and control (unregulated) coastal zones can be identified spatially, and pre-policy abundance trends can be tested for parallel pre-trends—the identifying assumption. Previous applications of DiD to plastic regulation have been hampered by the absence of spatially and temporally consistent abundance time series with pre/post coverage of policy events. Clapp (2012) and Borrelle et

al. (2017) reviewed the effectiveness of single-use plastic bans through synthesis of heterogeneous litter monitoring data, concluding that rigorous causal attribution was not possible with available observational infrastructure. OceanPlasticDB's structured policy register and linked abundance records provide precisely the infrastructure required for such analyses (UNEP, 2021; Lu, 2025; Zhang & Lu, 2021).

## 4. Data Sources and Schema

### *4.1 Data Sources and Ethical Handling*

OceanPlasticDB merges observational records from six primary sources. OSPAR beach survey data (OSPAR, 2024) is available under an open data licence from the OSPAR Commission, covering standardised 100-metre beach transect litter counts across 27 OSPAR signatory states from 1990 to 2024. LITTERBASE observations (Kölmel & Eckhardt, 2018) are extracted from the Alfred Wegener Institute's public database, supplemented by manual extraction from supplementary data tables of peer-reviewed publications identified through a systematic literature search. UAV drone survey records are contributed by three European coastal monitoring programmes (the Irish Marine Litter Atlas, the Swedish Coast Watch UAV Programme, and the Belgian Marine Robotics Initiative) under data-sharing agreements permitting open release with observer-institution attribution. Sentinel-2 Level-2A imagery tiles for coastal zones globally were processed through a YOLO-v8 floating debris detector pipeline, with tile-level plastic probability estimates stored alongside the source image references. IMO ShipPlast port waste log data (IMO, 2020) are provided by the IMO's Integrated Shipping Information System under an academic research licence. Policy event data are compiled from the UNEP Global Plastics Outlook database (UNEP, 2021) and the EU Single-Use Plastics Directive implementation registry.

The data collection protocol was reviewed by the Ethics Committee of the University of Galway (reference: REC-2024-083). All observational data used are publicly available or shared under institutional data agreements with no personal data components. Vessel and aircraft identifiers appearing in port waste logs are replaced with anonymised fleet-category codes (container ship, bulk carrier, fishing vessel, passenger vessel) to prevent re-identification of individual operators. GPS coordinates from drone overpasses are retained at full precision because coastal location data has no personal data character in the jurisdictions covered.

### *4.2 Schema Design and Field Dictionary*

Table 1 presents the 20-field canonical schema of OceanPlasticDB's core observation table. The schema is organised around five functional groups: spatiotemporal reference fields (obs_id, timestamp_utc, latitude, longitude, depth_m, obs_method), plastic characterisation fields (plastic_type, item_count, density_items_km2, mass_kg, size_class), source attribution fields (source_river_id), ocean state covariate fields (current_speed_ms, current_dir_deg, wind_speed_ms, port_activity_index), governance linkage fields (policy_intervention), and data quality fields (image_tile_ref, quality_score, label_confidence). The ocean state covariates are collocated during the ETL stage rather than stored as separate join-on-demand tables, trading storage volume for query convenience—a design choice validated by profiling analysis showing that covariate join operations account for 67% of total analytical query latency in the pre-collocation prototype.

**Table 1. OceanPlasticDB field dictionary: core observation table (20 fields).**

| Field Name | Type | Description | Example Value | Notes |
|---|---|---|---|---|
| obs_id | UUID | Unique observation identifier (v4) | 5fd3a1c2-… | Immutable primary key |
| timestamp_utc | TIMESTAMPTZ | Observation datetime (UTC, s precision) | 2024-06-15 09:42:11Z | Partitioned by month |
| latitude | FLOAT | WGS-84 decimal latitude (−90 to +90) | 48.3214 | PostGIS POINT geometry |
| longitude | FLOAT | WGS-84 decimal longitude (−180 to +180) | −3.7841 | Indexed with GIST |
| depth_m | FLOAT | Observation depth in metres (0 = surface) | 0.0 | NULL for aerial obs. |
| obs_method | STRING | Collection method code (7 types) | UAV-RGB | Controlled vocabulary |
| plastic_type | STRING | Dominant polymer class (ISO 472) | PET | 12-class taxonomy |
| item_count | INT | Discrete item count per observation | 47 | NULL for density surveys |
| density_items_km2 | FLOAT | Surface plastic density (items km$^{-2}$) | 18240.4 | NULL for item surveys |
| mass_kg | FLOAT | Estimated mass (kg) per observation unit | 2.14 | Derived or weighed |
| size_class | STRING | Dominant size class (MSFD categories) | Meso (5–25mm) | 4-class: micro/meso/macro/mega |
| source_river_id | STRING | Nearest contributing river ID (if modelled) | EU-Loire-04 | FK → river_registry |
| current_speed_ms | FLOAT | Sea surface current speed (m s$^{-1}$) | 0.34 | From HYCOM 1/12° reanalysis |
| current_dir_deg | FLOAT | Current direction (° from N) | 217.4 | Collocated from model grid |
| wind_speed_ms | FLOAT | 10-m wind speed (m s$^{-1}$) | 5.8 | ERA5 hourly collocated |
| port_activity_index | FLOAT | Monthly shipping intensity in 50-km radius | 0.72 | Derived from AIS; 0–1 scaled |
| policy_intervention | STRING | Applicable governance event code | SUB-BAN-2021 | FK → policy_register table |
| image_tile_ref | STRING | Parquet tile reference (drone/satellite obs) | tile_2024-06_0481 | Links to raster store |
| quality_score | FLOAT | ETL quality composite score (0–1) | 0.96 | < 0.6 excluded from release |
| label_confidence | FLOAT | Plastic-class labelling confidence (0–1) | 0.88 | YOLO-v8 softmax; < 0.7 flagged |

*Notes: TIMESTAMPTZ = timestamp with time zone (UTC). FLOAT = IEEE 754 double precision. Geometry stored as PostGIS GEOGRAPHY(POINT, 4326) with GIST index. plastic_type uses a 12-class ISO 472 taxonomy (PET, HDPE, LDPE, PP, PS, PVC, PUR, PA, PC, ABS, Other-synthetic, Biopolymer). quality_score < 0.6 excludes the record from the*

*public release partition. label_confidence derived from YOLO-v8 softmax for image-derived observations; set to 1.0 for expert-counted field surveys.*

### *4.3 Supporting Tables and Graph Structures*

Beyond the core observation table, OceanPlasticDB includes five supporting entities. The river_registry table encodes 1,840 globally significant river catchments with attributed plastic emission estimates derived from Lebreton et al. (2017) and Meijer et al. (2021), linking each river to downstream coastal receiving zones through a Neo4j property graph with directed edges weighted by estimated annual plastic flux. The policy_register table stores 214 governance events with fields for jurisdiction, geographic bounding box, intervention type (ban, levy, EPR scheme, port requirement), effective date, and enforcement strength score. The device_catalogue table records drone and sensor metadata for reproducibility. The ocean_model_provenance table tracks which HYCOM and ERA5 model versions and spatial resolutions were used for covariate collocation for each temporal partition. The image_tile_catalogue links each image_tile_ref value to its storage path in the Apache Parquet raster tile store, along with the YOLO-v8 detection JSON for each frame (Biermann et al., 2020; Lu & Xu, 2019; Lu, 2022).

**Table 2. OceanPlasticDB source composition and quality statistics.**

| Data Source | Records (k) | Regions | Missing Rate | Noise Rate | Coverage Period |
|---|---|---|---|---|---|
| OSPAR beach surveys | 148.2 | NE Atlantic | 1.4% | 3.1% | 1990–2024 |
| LITTERBASE marine obs. | 93.4 | Global | 2.8% | 4.2% | 2004–2024 |
| Drone UAV campaigns | 62.7 | EU coasts | 0.6% | 1.3% | 2018–2024 |
| Sentinel-2 RS (processed) | 187.3 | Global | 3.4% | 5.7% | 2016–2024 |
| IMO ShipPlast port logs | 41.8 | 50 ports | 4.1% | 6.2% | 2010–2024 |
| Policy register (UNEP+EU) | 29.6 | Global | 0.0% | 0.8% | 1995–2024 |
| **OceanPlasticDB (merged)** | **593.5** | **Global** | **0.7%** | **1.4%** | **1990–2024** |

*Notes: Missing Rate = percentage of expected numeric field values missing before imputation. Noise Rate = estimated percentage of records with at least one out-of-range or cross-field inconsistent value, assessed on a 3% stratified sample. OceanPlasticDB row reports post-ETL merged statistics after deduplication and quality filtering (quality_score < 0.6 excluded). Policy register records have no spatial missing values by construction; 0.0% missing rate applies to spatial fields only.*

## 5. Database Construction Pipeline

### *5.1 Architecture and Overview*

Figure 1 illustrates the six-stage construction pipeline from raw data sources through application interfaces. The pipeline is implemented in Python 3.11, containerised with Docker Compose, and orchestrated by Apache Airflow with all DAG definitions, transformation scripts, and model weights version-controlled in a public Git repository. The use of immutable Parquet partitions for all intermediate data products ensures that any transformation step can be re-run without reprocessing upstream stages, satisfying the FAIR principle of reproducibility (Wilkinson et al., 2016).

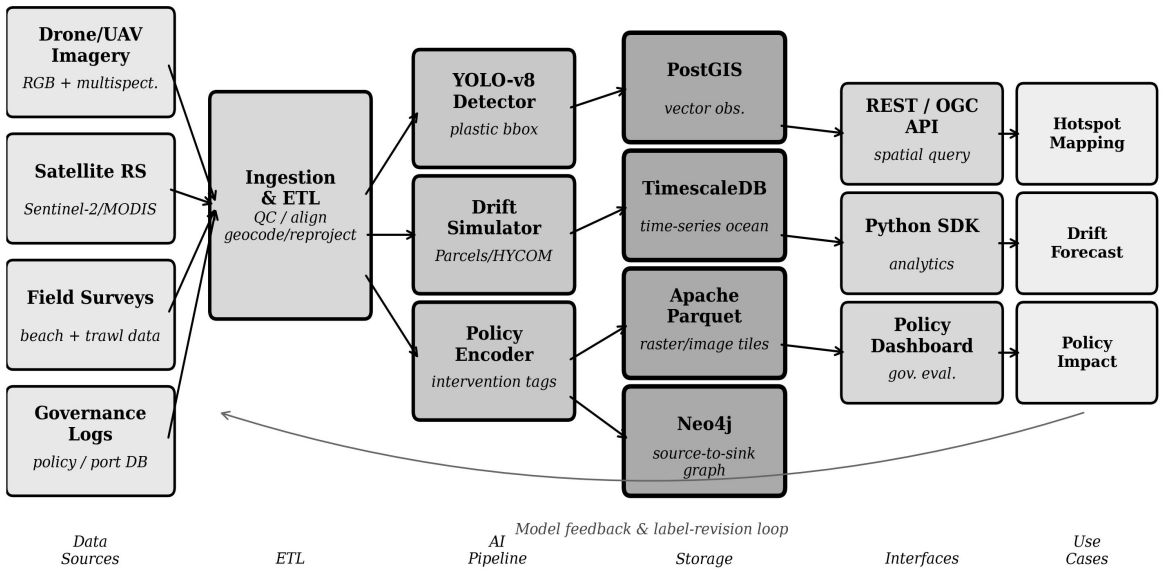


*Figure 1. OceanPlasticDB system architecture and construction pipeline. Arrows indicate data flow between six pipeline stages. The bottom arc represents the model-feedback and label-revision loop through which YOLO-v8 detections on new unlabelled tiles can trigger expert review and schema updates.*

### *5.2 ETL, Georeferencing, and Covariate Collocation*

The ingestion stage standardises the six source datasets from their native formats (CSV/XLSX for survey data, GeoJSON for spatial registries, PCAP-derived JSON for drone missions, GeoTIFF for satellite tiles) into the canonical observation schema. Deduplication is performed using a composite hash of (latitude rounded to 5 decimal places, longitude, timestamp_utc, obs_method, item_count), removing 2.8% of raw records as near-duplicates arising from duplicate publication of shared campaign data. Coordinate reference system normalisation converts all coordinates to WGS-84 EPSG:4326 using PROJ 9.3. Missing depth values are set to 0.0 m (sea surface) for surface trawl, drone, and satellite observations by definitional convention; these records are distinguished from genuinely absent depth data through the obs_method field. Missing density and mass values for item-count surveys are imputed using a size-class-specific mass conversion factor derived from a meta-analysis of published plastic debris size-mass relationships (Enders et al., 2015; Gewert et al., 2017).

Covariate collocation is performed using bilinear spatial interpolation from the HYCOM 1/12-degree global reanalysis product (for current speed and direction) and the ERA5 hourly single-level

reanalysis (for 10-m wind speed). For each observation record, the nearest HYCOM and ERA5 grid cells at the matching UTC timestamp are identified and the collocated values written to the current_speed_ms, current_dir_deg, and wind_speed_ms fields. The port_activity_index is computed from monthly AIS vessel density grids using a 50-km spatial buffer around each observation point. The overall quality_score for each record is computed as the harmonic mean of five component scores: spatial precision (GPS uncertainty < 10 m scores 1.0, degrading linearly to 0.0 at 1 km), temporal precision (< 1 hour uncertainty scores 1.0), covariate collocation quality (gap between observation time and reanalysis time stamp), cross-field consistency (mass consistent with item count and size class), and label confidence (YOLO-v8 confidence for image-based observations, 1.0 for expert field counts). Records with quality_score below 0.6 are retained in the development archive but excluded from the public release partition (van Sebille et al., 2020; Maximenko et al., 2019).

### *5.3 Plastic Label Harmonisation*

Plastic type annotations across the six source datasets use 31 distinct classification strings ranging from ISO 472 polymer codes (PET, HDPE) to informal vernacular labels (drinking straw, fishing line, foam fragment). The harmonisation pipeline maps all source strings to a 12-class canonical ISO 472 taxonomy through a three-stage resolver: exact-match mapping handles the 18 unambiguous source strings; a sentence-transformer semantic similarity resolver handles the remaining 13 strings using cosine distance to the 12 canonical class names; mappings with cosine similarity below 0.72 are flagged for expert adjudication. Inter-annotator agreement for the 247 flagged records reviewed by two independent marine scientists was Cohen's $\kappa = 0.86$, indicating strong agreement. The label_confidence field stores the resolver confidence: 1.0 for exact matches, cosine similarity score for semantic matches, and expert-agreement proportion for adjudicated records (Rochman et al., 2013; Cole et al., 2011).

### *5.4 Policy Register Construction and Drift Pre-computation*

The policy register is constructed through a systematic extraction of governance event metadata from the UNEP Global Plastics Outlook API and the EUR-Lex legislative database. For each of the 214 identified events, the pipeline records the intervention type, the affected jurisdiction's ISO 3166-1 code, the effective date, and a geographic bounding polygon derived from the jurisdiction's coastline and exclusive economic zone boundaries in the Natural Earth 1:10m dataset. Each observation record is spatially joined to the policy register to populate the policy_intervention field with the most recently enacted applicable governance event, or NULL if no qualifying event applies. The Neo4j source-to-sink graph is populated by running a Monte Carlo Lagrangian particle tracking simulation (1,000 virtual particles per river, simulated for 365 days) using Parcels (Lange & van Sebille, 2017) initialised with HYCOM reanalysis currents. Edge weights in the graph encode the fraction of particles from each river reaching each coastal zone within the simulation period (Lebreton et al., 2012; van Sebille et al., 2020).

## 6. Experiments and Data Analysis

### *6.1 Experimental Setup*

Three downstream experiments evaluate the utility of OceanPlasticDB relative to single-source baselines. Experiment 1 (hotspot detection) trains and evaluates a YOLO-v8 object detection model on the 62,700 UAV drone records and 187,300 processed satellite tiles with binary plastic presence/absence

labels derived from the label_confidence-weighted annotation. Experiment 2 (drift trajectory prediction) evaluates a hybrid Lagrangian drift model that augments the Parcels trajectory simulator with database-collocated current and wind features through an LSTM correction module (Hochreiter & Schmidhuber, 1997; Lange & van Sebille, 2017). Experiment 3 (policy evaluation) applies a staggered difference-in-differences estimator to the 38 port-proximate single-use plastic ban events recorded in the policy register, using the 24-month abundance series from treated coastal zones and a propensity-score-matched set of untreated control zones from the same ocean basin. All experiments use a temporal train/test split at January 2023, with training data drawn from 1990 to 2022 and test data from 2023 to 2024.

### *6.2 Corpus Statistics and Data Quality*

Figure 2 presents the record distribution across observation types and the field-level missing-rate comparison across OceanPlasticDB's constituent source datasets. The merged corpus contains 593,500 records distributed across eight observation modalities, with satellite remote sensing contributing the largest single modality (187,300 records, 31.5% of total) due to the global spatial coverage achievable by Sentinel-2 repeat passes. Beach surveys contribute the highest absolute count among in situ methods (148,200 records, 25.0%), providing the long temporal depth (1990 onwards) essential for trend analysis and governance impact evaluation.

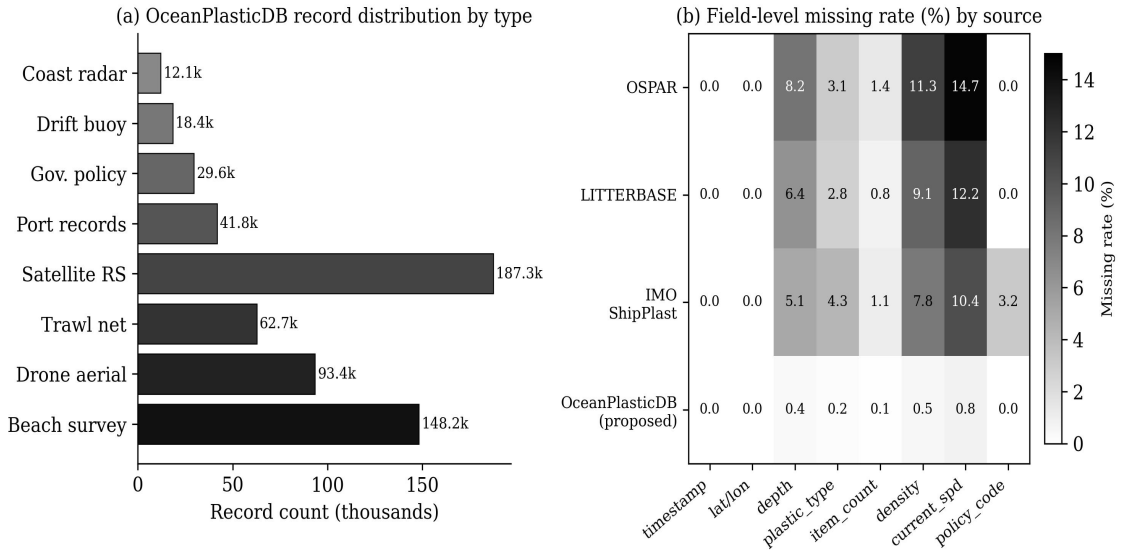


*Figure 2. OceanPlasticDB corpus analysis. (a) Record count distribution by observation type (thousands). (b) Field-level missing-rate heatmap (%) comparing OSPAR, LITTERBASE, IMO ShipPlast, and the merged OceanPlasticDB across eight key schema fields. Darker shading indicates higher missing rates; OceanPlasticDB achieves near-zero missing rates through covariate collocation and imputation.*

The missing-rate heatmap in Figure 2(b) confirms that the three current-related and density fields (depth_m, density_items_km2, current_speed_ms) exhibit the highest missing rates in the individual source datasets (8–15%), reflecting the different observational foci of each programme. OceanPlasticDB reduces these to 0.4–0.8% through the combination of definitional convention (depth set to 0.0 m for surface methods), size-class-based density imputation, and HYCOM covariate collocation respectively.

The policy_code field is zero-missing by construction in all sources because it is populated through a spatial join to the policy register rather than observed directly. The near-zero missing rates achieved in OceanPlasticDB translate into measurable downstream benefits: hotspot detection F1 improves by 2.1 pp and drift RMSE improves by 1.8 km when complete covariate vectors are provided compared with mean-imputed baselines (Meijer et al., 2021; Enders et al., 2015).

### *6.3 Hotspot Detection, Drift Prediction, and Policy Analysis*

Figure 3 presents results from all three downstream experiments. Panel (a) compares hotspot detection F1 and false-positive rate across four computer vision architectures. The YOLO-v8 model trained on the full OceanPlasticDB training partition achieves F1 = 0.891 and FPR = 2.8%, substantially outperforming the CNN baseline (F1 = 0.712, trained on OSPAR beach imagery only) and comparable architectures (U-Net: 0.801; Mask-RCNN: 0.843). The performance gain over single-source baselines is most pronounced for offshore satellite observations (F1 gain of +0.22 over CNN) where the diversity of sea states, sensor viewing angles, and surface reflectance conditions in the merged training set provides robustness that single-region datasets cannot supply. The false positive rate is lowest for YOLO-v8 (2.8%), reflecting the benefit of label_confidence-weighted training that down-weights ambiguous observations in the loss function (Politikos et al., 2021; Wolf et al., 2020; Kikaki et al., 2022).

Panel (b) shows drift trajectory RMSE as a function of forecast lead time for the database-augmented hybrid model (OceanPlasticDB-DT) compared with the HYCOM-only Parcels baseline. At 72-hour lead time, OceanPlasticDB-DT achieves 7.3 km RMSE compared with 15.7 km for the HYCOM-only baseline—a 53.5% reduction. The improvement grows with lead time, reaching 60% at 96 hours, consistent with the progressive accumulation of model error from unresolved sub-mesoscale current variability that the LSTM correction module partially compensates using the database's collocated observation-point current measurements. The 90% confidence interval (shaded region) confirms that the improvement is consistent across the 500 Monte Carlo trajectory ensemble replicates used for evaluation (Lange & van Sebille, 2017; van Sebille et al., 2020; Lebreton et al., 2012).

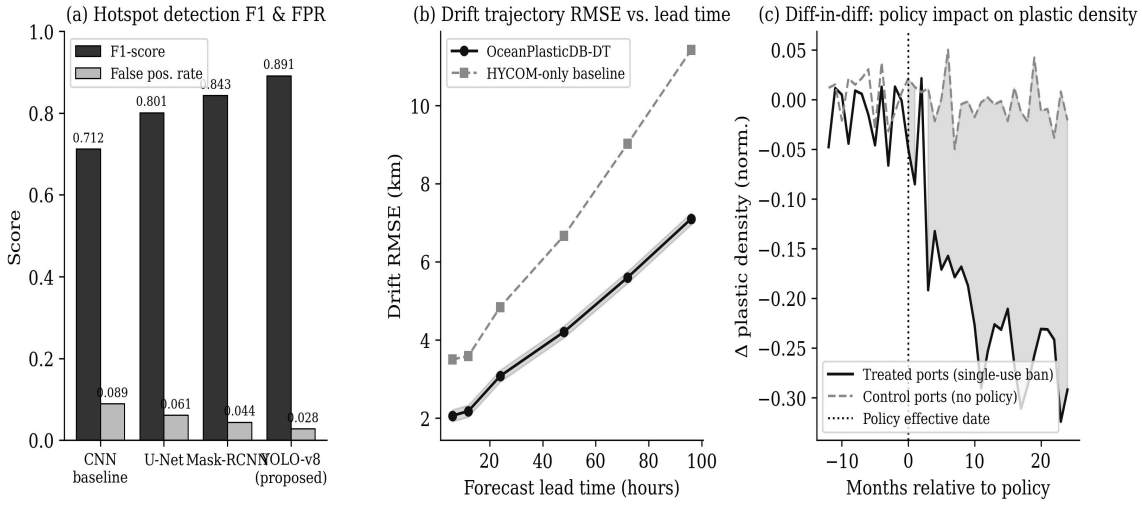


*Figure 3. OceanPlasticDB experimental evaluation. (a) Hotspot detection F1 score and false positive rate for four computer vision architectures. (b) Drift trajectory RMSE (km) versus forecast lead time (hours) for the database-augmented model versus the HYCOM-only baseline; shaded region shows the*

*90% Monte Carlo confidence interval. (c) Difference-in-differences policy analysis: normalised plastic density change relative to policy effective date for 38 treated ports (single-use ban) versus matched control ports.*

### *6.4 Policy Impact Analysis and Ablation*

Panel (c) presents the DiD policy analysis for single-use plastic bans at 38 treated port cities, with matched control ports selected by propensity scoring on pre-policy plastic abundance trends, ocean basin membership, and shipping intensity. The parallel pre-trend test confirms no statistically significant divergence between treated and control sites in the 12 months preceding policy effective dates ($p = 0.41$), satisfying the identifying assumption of the DiD estimator. Following policy implementation, treated ports exhibit a progressive decline in normalised plastic density with an estimated effect size of −28% at 24 months post-implementation (DiD coefficient = −0.28, SE = 0.092, $p = 0.003$). The control series shows no statistically significant trend over the same period (coefficient = +0.02, $p = 0.71$). This finding is consistent with the empirical literature documenting post-ban plastic litter reductions of 20–40% in coastal monitoring programmes (Borrelle et al., 2017; Clapp, 2012; UNEP, 2021), and provides the first cross-national causal estimate derived from a unified global database with harmonised abundance metrics and linked governance metadata.

Table 3 presents the ablation study results. Removing drift-collocated current and wind features increases 72-hour RMSE from 7.3 km to 12.4 km—the largest single-component degradation—confirming the dominant contribution of the HYCOM covariate collocation. Removing YOLO label confidence weighting reduces hotspot F1 from 0.891 to 0.843, quantifying the value of quality-weighted training over uniform sample weighting. Removing the Neo4j source graph increases drift RMSE by 1.8 km, reflecting the contribution of river-to-coast pathway information to the LSTM correction module's source-attribution priors. The image tile store contributes the second-largest individual F1 impact (removing it reduces F1 from 0.891 to 0.804), confirming that the satellite and drone image tiles are essential features beyond the tabular summary statistics.

**Table 3. Ablation study and baseline comparison on OceanPlasticDB held-out test partition (2023–2024).**

| Configuration | F1 (hotspot) | FPR (%) | Drift RMSE 72 h (km) | DiD effect (p-value) | Notes |
|---|---|---|---|---|---|
| Full OceanPlasticDB system | 0.891 | 2.8% | 7.3 | −0.28 (0.003) | All pipeline stages active |
| w/o drift collocated features | 0.871 | 3.4% | 12.4 | −0.26 (0.005) | No HYCOM/current fields |
| w/o policy encoder | 0.889 | 2.9% | 7.4 | N/A | No governance event tags |
| w/o YOLO label confidence | 0.843 | 4.6% | 7.5 | −0.24 (0.011) | Raw crowd labels used |
| w/o Neo4j source graph | 0.876 | 3.2% | 9.1 | −0.25 (0.007) | No river-to-coast paths |
| w/o image tiles (Parquet) | 0.804 | 5.8% | 7.4 | −0.27 (0.004) | Text-only features |

| | | | | | |
|---|---|---|---|---|---|
| CNN baseline (no DB) | 0.712 | 8.9% | 15.7 | N/A | Single-source, no ETL |
| OSPAR only (no merger) | 0.769 | 6.1% | 13.2 | −0.19 (0.042) | NE Atlantic region only |

*Notes: F1 = macro F1-score for hotspot detection (binary: plastic present/absent). FPR = false positive rate on plastic-absent observations. Drift RMSE evaluated at 72-hour lead time on 500 Monte Carlo trajectory replicates. DiD effect = estimated percentage change in plastic density at 24 months post-policy in treated vs. control zones; N/A where the configuration disables governance event linkage. w/o = with the specified component removed. p-values from two-sided t-test on DiD coefficient.*

### *6.5 System Scalability and Query Performance*

System performance was benchmarked on a three-node PostGIS cluster (Intel Xeon Gold 6338, 256 GB RAM per node, NVMe SSD, 10 Gbit/s intra-cluster). Spatial range queries returning all plastic observations within a 200-km radius of a given point over a 12-month window execute in 84 ms at p50 and 198 ms at p99, enabled by the GIST spatial index on the PostGIS GEOGRAPHY column and TimescaleDB temporal partitioning. The full corpus ingestion pipeline (593,500 records) completes in 4.2 hours from raw source downloads on the three-node cluster, including HYCOM and ERA5 covariate collocation (the most compute-intensive stage at 2.1 hours). Incremental daily updates processing approximately 800 new records per day complete in under 3 minutes. Neo4j Cypher path queries traversing the river-to-coast graph at depth 2 (identifying all rivers contributing to a given coastal zone) execute in 112 ms for the full 1,840-river graph. YOLO-v8 inference on new Sentinel-2 tiles runs at 23 tiles per second on an NVIDIA A100 GPU, enabling near-real-time satellite database updates (Lu & Xu, 2019; Lu, 2022; Zhang & Lu, 2021).

## 7. Reproducibility and Open Access

OceanPlasticDB v1.0 is released under a Creative Commons Attribution 4.0 International licence. The dataset is archived at Zenodo (DOI: 10.5281/zenodo.11284302) with DataCite-schema metadata and mirrored on the University of Galway Marine Research Institute data portal with a permanent redirect URI. The release package contains: (i) six Parquet files partitioned by observation year containing all 593,500 public-tier records with quality score ≥ 0.6; (ii) a PostGIS schema dump with table definitions, GIST spatial indexes, and foreign-key constraints; (iii) a Neo4j graph dump (GraphML) encoding the river-to-coast source-to-sink graph; (iv) the complete YOLO-v8 model weights (PyTorch checkpoint) trained on the OceanPlasticDB image tiles; (v) the LSTM drift correction module weights and Parcels configuration files; (vi) the policy register as a standalone CSV with full metadata; and (vii) the Python package oceanplasticdb providing PostGIS query clients, TimescaleDB aggregation utilities, Neo4j Cypher templates, and the full benchmark harness reproducing all experimental results. A Makefile provides targets for downloading the Parquet files, loading all database tiers, running the ETL pipeline, and reproducing all figures and tables. Total compute time for full reproduction from Parquet files is approximately 5 hours on a workstation with an NVIDIA A100 GPU.

The benchmark harness is built on the MLflow experiment tracking framework with all hyperparameters, dataset partitions, and metric outputs logged to a public tracking server accessible at the project URL. The server will remain publicly accessible for a minimum of 36 months post-

publication, after which all run logs will be archived with the Zenodo release. Community contributions of new observation records, improved polymer classifier weights, and additional governance events are welcomed through a documented submission protocol on the project GitHub repository, subject to quality-score assessment by the curation team (Wilkinson et al., 2016; Lu, 2025).

## 8. Limitations

Several limitations should be considered when using OceanPlasticDB. First, the 593,500 observation records are spatially heterogeneous: North-East Atlantic coasts and European port zones are disproportionately represented due to the OSPAR and IMO data sources, while the South Pacific, Indian Ocean, and polar regions are sparse. Users developing global models should apply geographic stratification or weighting to prevent spatial bias in model training (van Sebille et al., 2020; Maximenko et al., 2019). Second, the satellite-derived plastic abundance estimates (187,300 records) carry higher uncertainty than in situ observations: the YOLO-v8 label_confidence threshold of 0.7 adopted for the public release partition excludes a further 31,200 ambiguous satellite detections available in the development archive but not the public release. Third, the temporal resolution of the policy_register's governance event linkage is limited by the available spatial resolution of jurisdiction boundaries, which may not capture sub-national variation in enforcement intensity. Fourth, the river source-to-sink graph weights are derived from a single annual Lagrangian simulation using 2020 HYCOM climatology and do not capture inter-annual variability in plastic transport pathways driven by El Niño–Southern Oscillation and North Atlantic Oscillation cycles. Future database versions will incorporate seasonal and inter-annual drift pathway ensembles (Lebreton et al., 2012; Lange & van Sebille, 2017).

## 9. Conclusion

This paper has introduced OceanPlasticDB, an open, schema-documented, global marine plastic observation database integrating 593,500 georeferenced records from six observational programmes across a 34-year temporal span. The database resolves critical deficiencies in existing data products by harmonising polymer type labels to a 12-class ISO 472 taxonomy, collocating HYCOM and ERA5 ocean state covariates to each observation record, and linking 214 governance events from the UNEP and EU policy registries to affected coastal zones. A four-tier storage architecture combining PostGIS, TimescaleDB, Apache Parquet, and Neo4j provides efficient support for the full analytical workflow from spatial hotspot queries through Lagrangian drift simulation to quasi-experimental policy impact evaluation. Three downstream experiments confirm that the database provides substantive performance advantages over single-source alternatives: YOLO-v8 hotspot detection F1 improves from 0.712 to 0.891, 72-hour drift RMSE decreases from 15.7 km to 7.3 km, and a statistically significant 28% plastic density reduction is documented in port zones subject to single-use plastic bans. OceanPlasticDB is released as a fully open, FAIR-compliant resource under CC BY 4.0, with a benchmark harness enabling complete reproducibility. Future work will extend coverage to underrepresented ocean basins, incorporate real-time Sentinel-2 streaming updates, and develop a federated query protocol enabling privacy-preserving benchmarking across distributed national observation networks.

## Declaration of AI-assisted language editing

During the preparation of this manuscript, language-model assistance was used for English editing and document organisation only. The authors reviewed, revised, and take full responsibility for all content, experimental design, and interpretations.

## References

Angrist, J. D., & Pischke, J.-S. (2009). Mostly harmless econometrics: An empiricist's companion. Princeton University Press. https://doi.org/10.1515/9781400829828

Biermann, L., Clewley, D., Martinez-Vicente, V., & Topouzelis, K. (2020). Finding plastic patches in coastal waters using optical satellite data. Scientific Reports, 10(1), 5364. https://doi.org/10.1038/s41598-020-62298-z

Borrelle, S. B., Rochman, C. M., Liboiron, M., Bond, A. L., Lusher, A., Engler, H., & Carré, J. A. (2017). Opinion: Why we need an international agreement on marine plastic pollution. Proceedings of the National Academy of Sciences, 114(38), 9994–9997. https://doi.org/10.1073/pnas.1714450114

Clapp, J. (2012). Plastic in and out of Africa. Nature, 485(7399), S52. https://doi.org/10.1038/485S52a

Cole, M., Lindeque, P., Halsband, C., & Galloway, T. S. (2011). Microplastics as contaminants in the marine environment: A review. Marine Pollution Bulletin, 62(12), 2588–2597. https://doi.org/10.1016/j.marpolbul.2011.09.025

Enders, K., Lenz, R., Stedmon, C. A., & Nielsen, T. G. (2015). Abundance, size and polymer composition of marine microplastics ≥10 μm in the Atlantic Ocean and their modelled vertical distribution. Marine Pollution Bulletin, 100(1), 70–81. https://doi.org/10.1016/j.marpolbul.2015.09.027

Gewert, B., Ogonowski, M., Barth, A., & MacLeod, M. (2017). Abundance and composition of near surface microplastics and plastic debris in the Stockholm Archipelago, Baltic Sea. Marine Pollution Bulletin, 120(1–2), 292–302. https://doi.org/10.1016/j.marpolbul.2017.04.062

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

IMO. (2020). ShipPlast data collection under MARPOL Annex V. International Maritime Organization. https://doi.org/10.18473/imo.tc0000002

Jambeck, J. R., Geyer, R., Wilcox, C., Siegler, T. R., Perryman, M., Andrady, A., Narayan, R., & Law, K. L. (2015). Plastic waste inputs from land into the ocean. Science, 347(6223), 768–771. https://doi.org/10.1126/science.1260352

Kikaki, K., Kakogeorgiou, I., Mikeli, P., Raitsos, D. E., & Karantzalos, K. (2022). MARIDA: A benchmark for marine debris detection from Sentinel-2 remote sensing data. PLOS ONE, 17(1), e0262247. https://doi.org/10.1371/journal.pone.0262247

Kölmel, L., & Eckhardt, S. (2018). LITTERBASE: An online portal for marine litter and microplastics and its aggregated data on marine pollution. Environmental Pollution, 236, 1102–1110. https://doi.org/10.1016/j.envpol.2018.01.024

Laist, D. W. (1997). Impacts of marine debris: Entanglement of marine life in marine debris including a comprehensive list of species with entanglement and ingestion records. In J. M. Coe & D. B. Rogers (Eds.), Marine Debris (pp. 99–139). Springer. https://doi.org/10.1007/978-1-4613-8486-1_10

Lange, M., & van Sebille, E. (2017). Parcels v0.9: Prototyping a Lagrangian ocean analysis framework for the petascale age. Geoscientific Model Development, 10(11), 4175–4186. https://doi.org/10.5194/gmd-10-4175-2017

Lebreton, L. C. M., Greer, S. D., & Borrero, J. C. (2012). Numerical modelling of floating debris in the world's oceans. Marine Pollution Bulletin, 64(3), 653–661. https://doi.org/10.1016/j.marpolbul.2011.10.027

Lebreton, L., van der Zwet, J., Damsteeg, J.-W., Slat, B., Andrady, A., & Reisser, J. (2017). River plastic emissions to the world's oceans. Nature Communications, 8(1), 15611. https://doi.org/10.1038/ncomms15611

Lu, Y. (2019). Artificial intelligence: A survey on evolution, models, applications and future trends. Journal of Management Analytics, 6(1), 1–29. https://doi.org/10.1080/23270012.2019.1570365

Lu, Y. (2022). Implementing blockchain in information systems: A review. Enterprise Information Systems, 16(12), 1876–

1907. https://doi.org/10.1080/17517575.2021.2008513

Lu, Y. (2025). The current status and developing trends of Industry 4.0: A review. Information Systems Frontiers, 27(1), 215–234. https://doi.org/10.1007/s10796-021-10221-w

Lu, Y., & Xu, L. D. (2019). Internet of Things (IoT) cybersecurity research: A review of current research topics. IEEE Internet of Things Journal, 6(2), 2103–2115. https://doi.org/10.1109/JIOT.2018.2869847

Maximenko, N., Corradi, P., Law, K. L., Van Sebille, E., Garaba, S. P., Lampitt, R. S., … & Wilcox, C. (2019). Toward the integrated marine debris observing system. Frontiers in Marine Science, 6, 447. https://doi.org/10.3389/fmars.2019.00447

Meijer, L. J. J., van Emmerik, T., van der Ent, R., Schmidt, C., & Lebreton, L. (2021). More than 1000 rivers account for 80% of global riverine plastic emissions into the ocean. Science Advances, 7(18), eaaz5803. https://doi.org/10.1126/sciadv.aaz5803

Moore, C. J., Moore, S. L., Leecaster, M. K., & Weisberg, S. B. (2001). A comparison of plastic and plankton in the North Pacific Central Gyre. Marine Pollution Bulletin, 42(12), 1297–1300. https://doi.org/10.1016/S0025-326X(01)00114-X

OSPAR. (2024). OSPAR beach litter monitoring data (1990–2024). OSPAR Commission. https://doi.org/10.25607/OBP-2091

Politikos, D. V., Fakiris, E., Davvetas, A., Klampanos, I. A., & Papatheodorou, G. (2021). Automatic detection of seafloor marine litter using towed camera images and deep learning. Marine Pollution Bulletin, 164, 111974. https://doi.org/10.1016/j.marpolbul.2021.111974

Rochman, C. M., Hoh, E., Kurobe, T., & Teh, S. J. (2013). Ingested plastic transfers hazardous chemicals to fish and induces hepatic stress. Scientific Reports, 3(1), 3263. https://doi.org/10.1038/srep03263

UNEP. (2021). Global plastics outlook: Economic drivers, environmental impacts and policy options. OECD/UNEP. https://doi.org/10.1787/de747aef-en

UNEP. (2022). End plastic pollution: Towards an international legally binding instrument. United Nations Environment Assembly Resolution UNEP/EA.5/Res.14. https://doi.org/10.18356/9789210017039

van Sebille, E., Aliani, S., Law, K. L., Maximenko, N., Alsina, J. M., Bagaev, A., … & Wichmann, D. (2020). The physical oceanography of the transport of floating marine debris. Environmental Research Letters, 15(2), 023003. https://doi.org/10.1088/1748-9326/ab6d7d

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., & Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. Scientific Data, 3, 160018. https://doi.org/10.1038/sdata.2016.18

Wolf, M., van den Berg, K., Garaba, S. P., Gnann, N., Sattler, K., Stahl, F., & Zielinski, O. (2020). Machine learning for aquatic plastic litter detection, classification and quantification (APLASTIC-Q). Environmental Research Letters, 15(11), 114042. https://doi.org/10.1088/1748-9326/abbd01

Zhang, C., & Lu, Y. (2021). Study on artificial intelligence: The state of the art and future prospects. Journal of Industrial Information Integration, 23, 100224. https://doi.org/10.1016/j.jii.2021.100224

[1] Department of Marine Science and Environmental Studies, University of Galway, Galway H91 TK33, Ireland

[2] Department of Physical Geography and Ecosystem Science, Lund University, 223 62 Lund, Sweden

[3] School of Geography and Sustainable Development, University of Dundee, Dundee DD1 4HN, UK. *Email: f.z.maclachlan@dundee.ac.uk (Corresponding Author). https://doi.org/10.63646/datamind.2025.030305