Normalization Layer Spec

The Normalization Layer Spec defines the transformation rules that every inbound telemetry signal must pass before it enters fusion. Normalization is not optional pre-processing; it is the contract boundary between raw telemetry and operationally useful data. Its four responsibilities — format alignment, timestamp normalization to UTC, and handling of missing, duplicate, and conflicting signals — must be declared once, versioned, and applied uniformly to every source_id registered in e5-telemetry-source-map. Any datum that fails normalization without a declared handling rule is quarantined, not silently dropped; its quarantine event is a provenance record that feeds downstream quality tracking.

1. Scope and Non-Negotiables

This spec governs all telemetry sources regardless of their confidence_band. A high source and a low source pass through the same normalization rules; the band difference affects fusion weighting and alert authority, not the cleaning obligation. The normalization layer may not:

- Impute a value for a missing field without declaring it as synthetic in the output record.
- Silently discard a duplicate without logging the discard event.
- Resolve a conflict between sources autonomously — conflict records are forwarded to the fusion layer with both values intact and flagged.
- Alter a raw measurement value to fit an expected range without retaining the original value in the provenance envelope.

2. Format Alignment Rules

Each telemetry source in e5-telemetry-source-map is assigned a normalization profile that maps its raw field names and units to the canonical schema. The canonical output record is:

```
normalized_record:
  source_id: <resolves to e5-telemetry-source-map row>
  measurement_type: <enum; see §2.1>
  value: <number or string per type>
  unit: <canonical unit per type>
  confidence_band: <inherited from source_id row; may be downgraded by §5>
  timestamp_utc: <ISO 8601 with Z suffix>
  raw_value: <original unmodified value; preserved for audit>
  raw_unit: <original unit before conversion>
  synthetic: <true|false; true if value was imputed>
  normalization_profile_version: <semver>
```

2.1 Canonical measurement types and units

measurement_type	canonical unit	typical raw variants
water-level	meters above NAVD88	feet, cm, mm, elevation codes
rainfall-rate	mm/hour	in/hr, mm/min, count per tip

measurement_type	canonical unit	typical raw variants
rainfall-accumulation	mm (rolling window declared in profile)	inches, cm
tide-height	meters MLLW	feet, cm, NAVD88 offset variants
soil-moisture-volumetric	m^3/m^3	% VWC, raw ADC counts
flow-rate	m^3/second	cfs, liters/min, velocity-based derived
ndwi-index	dimensionless (–1 to +1)	raw DN, percent
barometric-pressure	hPa	mbar, inHg
wind-speed	m/s	mph, knots

Unit conversions must be stored as static lookup constants — not derived inline — so they can be audited and versioned independently.

2.2 Field mapping per normalization profile

Each source's normalization profile is a YAML document versioned alongside this spec. A minimal profile entry:

```
profile_id: <source_id>-norm-v1
source_id: <resolves to e5-telemetry-source-map>
field_map:
  - raw_field: <raw JSON/CSV path>
    canonical_field: <canonical field name>
    unit_conversion: <formula or lookup key; "none" if already canonical>
timestamp_path: <dot-path to timestamp field in raw payload>
timestamp_format: <strftime pattern or "iso8601" or "unix-epoch-ms">
```

3. Timestamp Normalization

All timestamps are stored in UTC with the Z suffix regardless of source timezone. Sources that do not include a timezone offset are assumed to be in the operating jurisdiction's local time (default: America/Chicago for CST/CDT; override per source in the normalization profile). If the inferred timezone leads to a timestamp that is more than 10 minutes in the future relative to the ingest wall clock, the record is flagged clock-drift-suspect: true and emits a low-severity provenance event.

Ordering guarantee: The normalization layer does not guarantee arrival-order delivery; it guarantees event-time ordering within a single source_id stream. Cross-source temporal alignment is the responsibility of e5-fusion-logic-map, which uses a declared fusion window.

4. Missing Signal Handling

Missing signals fall into three categories:

Category	Definition	Handling
field-missing	Expected field absent from payload	Mark value: null, synthetic: false; quarantine if field is required by the normalization profile
feed-gap	No records received within 2× the expected emission_frequency window	Emit a feed-gap event to the pipeline monitoring channel; do NOT impute a value unless the profile explicitly declares an imputation rule
sensor-offline	Heartbeat absent beyond a configurable offline-threshold (default: 5× emission_frequency)	Mark source as status: offline in the source registry; demote confidence_band to low automatically; emit an alert to e5-alert-design-spec pipeline

Imputation is permitted only for rainfall-accumulation when a partial-window gap is shorter than 20% of the rolling window and the source's confidence_band is high or medium. Any imputed value carries synthetic: true.

5. Duplicate Signal Handling

A duplicate is any record from the same source_id with the same timestamp_utc as an already-ingested record.

1. Compare raw_value between the duplicate and the stored record.

2. If identical: discard silently but log a duplicate-discarded event with count and source.

3. If values differ: retain both records with a conflict_group_id (shared UUID v4), set duplicate_conflict: true on each, and forward the conflict group to the fusion layer. Do NOT pick a winner here.

High-frequency sources (emission_frequency ≤ sub-minute) may configure a deduplication window (default: 2 seconds) within which near-simultaneous records with the same value are treated as category-1 duplicates.

6. Conflicting Signal Handling

A conflict arises when two or more sources claiming authoritative: true for the same measurement_type emit materially different values at the same point in time. The normalization layer does not resolve conflicts; it documents them.

Conflict detection threshold per measurement type:

measurement_type	conflict threshold
water-level	> 0.05 m difference
rainfall-rate	> 2 mm/hr difference
tide-height	> 0.03 m difference

measurement_type	conflict threshold
soil-moisture-volumetric	> 0.05 m³/m³ difference
flow-rate	> 0.01 m³/s difference

When a conflict is detected, the normalization layer emits a conflict-group record containing both normalized values, both source_ids, both confidence_bands, and the magnitude of the delta. The fusion layer then applies its weighting rules (see e5-fusion-logic-map). Conflicts between a high-band and a low-band source never silently promote the low-band value.

7. Band Downgrade Triggers

The normalization layer inherits confidence_band from e5-telemetry-source-map but may downgrade (never upgrade) based on observed data quality:

Trigger	Downgrade action
feed-gap event	Downgrade by one notch for the duration of the gap
sensor-offline	Downgrade to low until heartbeat restored
clock-drift-suspect	Downgrade by one notch; alert to pipeline monitoring
Consecutive field-missing > 3 within 10× emission window	Downgrade by one notch; notify source owner

Downgrades are recorded in the normalized record's confidence_band field and in a provenance event. Re-upgrade to the base band requires a fresh last_validated attestation by the source owner.

8. Worked Example — Pond Level Normalization (illustrative)

Source sw-water-level-j1 (MQTT/TLS, sub-minute) emits:

```
{
  "node": "j1",
  "level_ft": 2.47,
  "ts": "2026-05-29T14:32:11",
  "battery_pct": 91
}
```

After normalization using profile sw-water-level-j1-norm-v1:

```
source_id: sw-water-level-j1
measurement_type: water-level
value: 0.753
unit: meters above NAVD88
confidence_band: medium          # inherited; pending calibration
timestamp_utc: "2026-05-29T19:32:11Z"   # CDT → UTC (+5h)
raw_value: 2.47
raw_unit: feet
```

```
synthetic: false
normalization_profile_version: "1.0.0"
```

The conversion is 2.47 ft × 0.3048 = 0.753 m. The local timestamp 14:32:11 CDT is converted to 19:32:11Z. If the station had emitted 14:32:11 without a timezone tag and the profile declares America/Chicago, the same UTC conversion applies.

Now suppose a second record arrives 1.1 seconds later with level_ft: 2.47 and the same timestamp. The deduplication window (2 s) catches it; a duplicate-discarded event is logged; no second record enters the fusion layer.

Suppose instead the culvert flow sensor sw-flow-culvert-c4 (grade C, confidence_band: low) reads an unusually high value at the same moment. Normalization does not merge or suppress this; it forwards a conflict-group record to e5-fusion-logic-map with both values, both bands, and the delta flagged.

9. Normalization Profile Versioning

Normalization profiles are versioned independently of this spec. A change that affects the value field of any existing measurement type (unit conversion constant, field mapping, or imputation rule) is a MINOR version bump and requires a re-normalization sweep for any data in the fusion window. A change that adds a new measurement_type is also MINOR. Removing a measurement type or changing a canonical unit is MAJOR and requires a coordinated bump with e5-fusion-logic-map.