



Complete Guide –Water Quality Data Validation App

Version: v1.0
Date: 2025-09-05

1) Quick Start

1) Run:

`streamlit run Water_Validation_App.py`

- 2) In the “ Upload File” tab, upload your raw Excel file once.
- 3) Go to “ Run All & Exports” and click “Run All Steps”.
- 4) Download your outputs:
 - Final_Combined.xlsx
 - Final_Repaired.xlsx
 - The ZIP bundle containing all cleaned/annotated files + the two finals

Tip: You can also run each stage individually in its tab and download the per-stage cleaned/annotated files.

2) What each tab does

Upload File

- Loads the raw Excel and keeps it in memory for subsequent steps.
- Shows a 20-row preview.

GENERAL Validation

- Removes duplicate rows by identity key: Group/Affiliation + Site + Date + Time.
- Drops rows flagged by any column whose name contains “flag” and has values like yes/true/1/...
- Enforces “ ≥ 3 sites per group/watershed” and “ ≥ 10 events per site”.
- Validates standard ranges (pH, DO, temperature, salinity, conductivity, ...).
- Detects contextual outliers ($>3\sigma$ within the same site).
- Notes Sample Time if it falls within 12:00–16:00 (warning only).
- Flags “expired reagents”; requires a comment if issues exist.
- Outputs: cleaned_GENERAL.xlsx and annotated_GENERAL.xlsx

CORE Validation

- Sample depth must be 0.3 m or mid-depth.
- Total Depth = 0 only with Flow = 6 (else row is dropped).
- DO measured twice; difference ≤ 0.5 mg/L (if greater \Rightarrow Note + round both to 0.1).
- Secchi: two significant figures and must not exceed total depth.
- Conductivity calibration: within $\pm 20\%$ of Standard Value.

- Pre/Post calibration times within ± 24 h of the sample time.
- Conductivity formatting: $>100 \Rightarrow$ up to 3 integer digits; $<100 \Rightarrow$ integer only.
- Round pH/Water Temp to 0.1; standardize Salinity display (" < 2.0 " or 0.1 rounding).
- Outputs: cleaned_CORE.xlsx and annotated_CORE.xlsx

3 ECOLI Validation

- Incubation temperature 30–36°C and time 28–31h.
- Colonies < 200 ; Field Blank must show no growth.
- E. coli = 0 \Rightarrow set to NaN.
- Two-step rounding for E. coli Average: nearest integer \rightarrow 2 significant figures.
- CFU formula check: $CFU = (Colonies \times Dilution \times 100) / Volume$.
- Outputs: cleaned_ECOLI.xlsx and annotated_ECOLI.xlsx

4 ADVANCED Validation

- Units: Nitrate/Phosphate = mg/L (or ppm), Turbidity = NTU, Streamflow/Discharge = ft²/sec.
- Discharge formatting: $<10 \Rightarrow$ one decimal; $\geq 10 \Rightarrow$ integer (changes logged).
- Outputs: cleaned_ADVANCED.xlsx and annotated_ADVANCED.xlsx

5 RIPARIAN Validation

- "Bank Evaluated" must be filled.
- Indicators must be present or justified in Comments; otherwise a Note is added.
- If multiple indicators are missing, a collective "Indicators incomplete" Note is added.
- Standardizes "Yes" for the site-image column.
- Outputs: cleaned_RIPARIAN.xlsx and annotated_RIPARIAN.xlsx

Run All & Exports

Runs all five stages end-to-end and saves every cleaned/annotated file.

Produces two final datasets:

- 1) Final_Combined.xlsx (all stage notes/changes merged)
- 2) Final_Repaired.xlsx (deterministic, safe value fixes applied)

Offers a ZIP bundle with all outputs.

Cleaning Guide

Download button for the "Validation Rules for Parameters" PDF (if placed next to the app).

3) Output types & when to use them

A. cleaned_*.xlsx

“Analysis-ready” subset: worst rows are removed; invalid/out-of-range/outlier values become NaN.
Use for maps, models, and statistics that must avoid bad values.

B. annotated_*.xlsx

The data plus explanatory columns (Notes/ChangeNotes and helper columns).
Use for auditing and understanding why values were removed/naïved/rounded.

C. Final_Combined.xlsx

The final cleaned dataset with ALL stage Notes/ChangeNotes merged per row
(GENERAL/CORE/ECOLI/ADVANCED/RIPARIAN).
Key columns: *_Notes, *_Changes, All_Notes, All_ChangeNotes.
Use for management reporting and one-glance diagnosis per row.

D. Final_Repaired.xlsx

The final dataset where deterministic, safe fixes are applied to the values:

- pH/Water Temp \Rightarrow rounded to 0.1
- DO \Rightarrow mean(DO1, DO2) if both present, rounded to 0.1
- Salinity \Rightarrow standardized display
- E. coli Average \Rightarrow nearest-integer then to 2 significant figures
- CFU \Rightarrow recomputed and replaced when inputs exist
- Conductivity column header unified: (?S/cm) \rightarrow (μ S/cm)
- Nitrate/Phosphate unit “ppm” \Rightarrow “mg/L” (no numeric change in freshwater)
- Duplicate rows removed by identity key

All changes are logged in Repaired_ChangeLog.

Use when you want a ready-to-consume, consistent dataset.

E. ZIP bundle

Contains all cleaned/annotated files + Final_Combined + Final_Repaired for archiving/sharing.

4) Final_Combined vs Final_Repaired

- Final_Combined: transparency-first — every stage’s Notes/Changes are visible per row.
- Final_Repaired: usability-first — safe value fixes are applied and documented in Repaired_ChangeLog.
- We do NOT “clamp” out-of-range values to the boundaries; those remain NaN unless you opt into a policy change.

5) Tips & Troubleshooting

- Conductivity column name: the app accepts both “Conductivity (μS/cm)” and “Conductivity (?S/cm)”; the repaired output unifies to μS/cm.
- Flag columns: any column whose name contains “flag” and values like yes/true/1/... marks the row for removal in cleaned outputs.
- Sample Time 12–16: not removed; a warning Note is added.
- DO titrations: if one column is missing, a “Missing DO titration column” Note is added. If both present and $\Delta > 0.5$ mg/L \Rightarrow Note + round to 0.1.
- If a stage’s cleaned output becomes empty, later stages fall back to the nearest non-empty cleaned source; the finals are built from the last non-empty cleaned dataset (priority: RIPARIAN \rightarrow ADVANCED \rightarrow GENERAL).
- Thresholds you can customize:
 - Standard ranges in run_general (dict: standard_ranges)
 - DO difference (≈ 0.5) in run_core
 - Incubation temperature/time ranges in run_ecoli
 - Unit/Discharge rules in run_adv
 - Riparian indicator list in run_rip

6) Mini Glossary

- cleaned: analysis-ready rows; worst rows removed; invalid cells as NaN.
- annotated: same data plus Notes/ChangeNotes columns.
- Final_Combined: cleaned base + all stage Notes/Changes per row.
- Final_Repaired: final data with safe value fixes applied; changes logged.
- 2SF: Two Significant Figures.
- CFU: Colony Forming Units per 100mL.

7) FAQ

- **Why fewer rows in cleaned than raw?**

Critical bad rows are removed (invalid date/time, all-core missing, depth/flow mismatch, flagged rows, etc.).

- **Which final should I publish?**

If transparency matters: Final_Combined. If you want ready-to-use values: Final_Repaired.

- **Can we clamp out-of-range values to min/max?**

Not by default (scientific/audit reasons). We can add it as an optional policy if you decide to.