Validation Rules for Parameters – GENERAL Stage

Row-Level Deletion Conditions

(→ These rows will NOT be included in `cleaned_GENERAL.xlsx`)

Condition	Explanation	Note Added
◆ Fewer than 3 sites in a watershed	Checked using `Group or Affiliation` + `Site ID: Site Name`	`Less than 3 sites in watershed;`
◆ Fewer than 10 events per site	Checked by counting unique `Sample Date` per site	`Fewer than 10 events;`
◆ Invalid or missing sample date	Checked via `Sample Date` column	`Missing or invalid Sample Date;`
◆ Unparsable sample time	Checked by testing if `Sample Time Final Format` can be parsed	`Unparsable Sample Time;`
◆ All core parameters missing or zero	If all the following are zero/NaN:	
• `pH (standard units)`		
• `Dissolved Oxygen (mg/L) Average`		
• `Water Temperature (° C)`		
• `Conductivity (?S/cm)`		
• `Salinity (ppt)`	`All core parameters missing or invalid;`	

Cell-Level Deletion Conditions

*(→ Affected **cells** are replaced with `NaN`, but the row is retained)*

Out-of-Range Values (Replaced with NaN)

For the following parameters, if values fall outside defined limits:

Parameter	Valid Range	Error Note
`pH (standard units)`	6.5 – 9.0	`pH out of range [6.5–9.0];`
`Dissolved Oxygen (mg/L) Average`	5.0 – 14.0	`DO out of range [5.0–14.0];`
`Conductivity (?S/cm)`	50 – 1500	`Conductivity out of range [50–1500];`
`Salinity (ppt)`	0 – 35	`Salinity out of range [0–35];`
`Water Temperature (° C)`	0 – 35	`Temp out of range [0–35];`

Parameter	Valid Range	Error Note
Air Temperature (° C)	-10 – 50	Air Temp out of range [-10–50];
Turbidity	0 – 1000	Turbidity out of range [0–1000];
E. Coli Average	1 – 235	E. Coli out of range [1–235];
Secchi Disk Transparency - Average	0.2 – 5	Secchi out of range [0.2–5];
Nitrate-Nitrogen VALUE (ppm or mg/L)	0 – 10	Nitrate out of range [0–10];
Orthophosphate	0 – 0.5	Orthophosphate out of range [0–0.5];
DO (%)	80 – 120	DO % out of range [80–120];
Total Phosphorus (mg/L)	0 – 0.05	TP out of range [0–0.05];

Contextual Outliers (Z-Score > ±3 within Site)

If a value is more than ±3 standard deviations from the site mean:
Note added: `{Parameter} is a contextual outlier (>3 std);`
Cell value is set to `NaN`

Expired Reagents Used

If the `"Chemical Reagents Used"` column contains the word `"expired"`:
Note added: `Expired reagents used;`
Cell value is set to `NaN`

Missing Comments for Problematic Values

If a flagged value is found but the `"Comments"` field is blank:
Note added: `No explanation in Comments;`

Strip the Words "valid"/"invalid"

If any cell contains the words `"valid"` or `"invalid"`, they are removed.
Note added in `TransformNotes`: `Removed 'valid/invalid';`

CORE Validation

Path: `with tabs[2]`
This tab handles the validation of key measurement parameters related to **depth**, **dissolved oxygen (DO)**, **transparency**, **calibration**, and other core metrics.

Outputs Generated

- `cleaned_CORE.xlsx` – Contains only valid rows (no deletion-triggering errors)
- `annotated_CORE.xlsx` – Contains all rows, with notes and change logs

Parameters Checked

Category	Columns Involved
Sampling Depth	`"Sample Depth (meters)"`, `"Total Depth (meters)"`
DO Titration	`"Dissolved Oxygen (mg/L) 1st titration"`, `"Dissolved Oxygen (mg/L) 2nd titration"`
Water Transparency	`"Secchi Disk Transparency - Average"`
Conductivity Calibration	`"Conductivity (?S/cm)"`, `"Post-Test Calibration Conductivity"`, `"Standard Value"`
Temperature and pH	`"Water Temperature (° C)"`, `"pH (standard units)"`
Miscellaneous	`"Time Spent Sampling/Traveling"`, `"Roundtrip Distance Traveled"`, `"Salinity (ppt)"`, `"Flow Severity"`

Row-Level Deletion Rule

Condition	Explanation	Note Added
`"Total Depth (meters)" == 0` AND `"Flow Severity" ≠ 6`	Indicates dry stream flag missing if depth is zero	`Zero Depth with non-dry flow;`

Cell-Level Notes and Flags (Row Kept)

Parameter	Condition	Note Added
Sampling Depth	Not equal to `0.3` **OR** not within ±0.05 of half total depth	`Sample Depth not 0.3m or mid-depth;`
DO Difference	`abs(DO1 - DO2) > 0.5`	`DO Difference > 0.5;`
Rounded DO	Round both DOs to 1 decimal place	Logged in `"CORE_ChangeNotes"`
Secchi Transparency	- Secchi > Total Depth	

- Secchi not to 2 significant figures | `Secchi > Depth;`, `Secchi not 2 significant figures;` |
  | Conductivity Calibration | Post-test calibration ≠ ±20% of standard | `Post-Test Calibration outside ±20% of standard;` |
  | Conductivity Format |
- Value >100 with more than 3 digits

- Value ≤100 but contains decimals | `Conductivity format error;` |
  | Salinity |
- If < 2.0, must be `< 2.0`
- Other values rounded to 1 decimal place | Logged in `"CORE_ChangeNotes"` |
  | pH and Water Temp | Rounded to 1 decimal place; auxiliary columns like `"pH Rounded"` created | Logged in `"CORE_ChangeNotes"` |
  | Time and Distance Fields | If values are non-numeric (letters or words) | `Time format not numeric;, Distance format not numeric;` |

Output Details

- **`cleaned_CORE.xlsx`**: Contains only rows without any row-deletion-level errors.
- **`annotated_CORE.xlsx`**:
  Contains **all rows**, and includes:
  - `"CORE_Notes"` – for listing validation issues
  - `"CORE_ChangeNotes"` – for logging changes
  - Additional helper columns (e.g., `"DO1 Rounded"`, `"pH Rounded"`) to show transformed values without replacing originals

ECOLI Validation

Path: `with tabs[3]`
This stage checks all fields related to **E. coli testing** — including incubation temperature/time, colony counts, blanks, value rounding, and CFU (Colony Forming Units) calculations.

Outputs Produced

- `cleaned_ECOLI.xlsx`: Only rows with **no major validation issues**
- `annotated_ECOLI.xlsx`: All rows + validation/change notes

Parameters Reviewed

Check Type	Parameters
Incubation Temperature	`"Incubation temperature is 33° C +/- 3° C"`
Incubation Time	`"Incubation time is between 28-31 hours"`
Colony Count	`"Sample 1: Colonies Counted"`, `"Sample 2: Colonies Counted"`
Field Blank	`"No colony growth on Field Blank"`
E. coli Value	`"E. Coli Average"`
CFU Calculation	Uses `Dilution` and `Sample Size` per sample
Rounding to 2 Sig Figs	Applied to `"E. Coli Average"`

Cell-Level Deletion Triggers (Value set to `NaN`)

Parameter	Condition	Note Added
Incubation Temperature	Value not in 30–36°C	`Incubation temperature not in 30–36°C range;`
Incubation Time	Value not in 28–31 hours	`Incubation time not in 28–31h range;`
Colony Count	Value > 200 (Sample 1 or 2)	`{col} > 200 colonies;`
E. coli Value	Value = 0	`E. coli = 0;`

Warnings Without Deletion (Notes only)

Condition	Parameter	Note Added
Field Blank Contains growth	`"No colony growth on Field Blank"` contains `"no"`, `"false"`, `"n"` (case-insensitive)	`Colony growth detected in field blank;`
CFU Formula Mismatch CFU = (Colonies × Dilution × 100) / Volume	Formula:	
If mismatch > ±10	`{prefix} CFU formula mismatch;`	
Rounding to 2 Significant Figures	Applied to `"E. Coli Average"`	`E. coli {orig} → {rounded} (rounded to 2 significant figures);` (in `"ECOLI_ChangeNotes"`)

Skip Unmeasured Columns

If an entire column is blank or filled with zeros, it is **excluded from validation**.
→ Note added in `"ECOLI_ChangeNotes"`:
`Skipped checks for unmeasured parameter: {col}`

Final Outputs Summary

`cleaned_ECOLI.xlsx`

- Includes **only** rows where `"ECOLI_ValidationNotes"` is **empty** (no errors)

`annotated_ECOLI.xlsx`

- Includes **all rows**
- Extra columns:
  - `"ECOLI_ValidationNotes"`: Describes errors/issues
  - `"ECOLI_ChangeNotes"`: Describes transformations (e.g., rounding)

- Calculated columns: `"E. Coli Rounded"`, `"E. Coli Rounded (2SF)"`

ADVANCED Validation

Path: `with tabs[4]`
This stage validates **advanced chemical and hydrological parameters**, focusing on:

- Parameter names and unit consistency
- Unit compatibility with parameter types
- Detection of all-zero columns
- Numeric formatting and precision (e.g., discharge rounding)

Outputs

- `cleaned_ADVANCED.xlsx`: Only rows **without any validation errors**
- `annotated_ADVANCED.xlsx`: All rows, with added **notes and change logs**

Step-by-Step Parameter Checks

1. Detect All-Zero or Empty Columns

Any numeric column entirely filled with zeros or empty values is skipped.
→ Added to `all_zero_cols`

Note in `"ADVANCED_ChangeNotes"`:
`Skipped checks for unmeasured parameter: {col};`

2. Parameter Label and Unit Checks

Type	Condition	Note Added
Phosphate	Column name contains `"Phosphate"` and `"Value"`, but **does not** contain `"mg/L"` or `"ppm"`	`Phosphate not labeled in mg/L`
Nitrate-Nitrogen	Same rule as above for `"Nitrate-Nitrogen"`	`Nitrate-Nitrogen not labeled in mg/L`
Turbidity	Column includes `"Turbidity"` and `"Result"`, but **not** `"NTU"` (e.g., shows `"JTU"`)	`Appears to be in JTU not NTU`

3. Discharge Format & Precision

Condition	Response	Note Added
Value < 10 and lacks 1 decimal	Round to 1 decimal	Logged in `"ADVANCED_ChangeNotes"`
Value ≥ 10 and has decimals	Round to integer	Logged in `"ADVANCED_ChangeNotes"`
Invalid or non-numeric value	Skip check	`Invalid or non-numeric discharge value;`

4. Unit-Characteristic Validation

Using:

- `"CharacteristicName"`
- `"ResultMeasure/MeasureUnitCode"`

Characteristic	Expected Unit	Note on Mismatch
Phosphate	`mg/L` or `ppm`	`Phosphate unit invalid: {unit}`
Nitrate-Nitrogen	`mg/L` or `ppm`	`Nitrate-Nitrogen unit invalid: {unit}`
Turbidity	`NTU`	`Turbidity unit should be NTU, found: {unit}`
Streamflow	`ft2/sec`	`Streamflow unit should be ft2/sec, found: {unit}`
Discharge	`ft2/sec`	`Discharge unit should be ft2/sec, found: {unit}`

5. Final Output Rules

`cleaned_ADVANCED.xlsx`

- Contains only rows with **no content in `"ADVANCED_ValidationNotes"`**

`annotated_ADVANCED.xlsx`

- Includes all data rows
- With additional columns:
  - `"ADVANCED_ValidationNotes"`: Describes all issues
  - `"ADVANCED_ChangeNotes"`: Logs all fixes or formatting changes

RIPARIAN Validation

Path: `with tabs[5]`

This final step validates **riparian assessment data**, focusing on qualitative or ranked indicators of vegetation, erosion, and site documentation.

Outputs

- `cleaned_RIPARIAN.xlsx`: Only rows **without critical validation issues**
- `annotated_RIPARIAN.xlsx`: All rows, with **notes and change logs**

Indicators Reviewed

Indicator Columns

```
Energy Dissipation
New Plant Colonization
Stabilizing Vegetation
Age Diversity
Species Diversity
Plant Vigor
Water Storage
Bank/Channel Erosion
Sediment Deposition
```

Step 1: Identify All-Zero or Empty Columns

- If a column contains **only zeros or NaNs**, it is excluded from validation.
- Note added in `"RIPARIAN_ChangeNotes"`:
  `Skipped checks for unmeasured parameter: {col}`

Step 2: Check `"Bank Evaluated"` Column

Condition	Explanation	Note Added
Empty or neutral (e.g., "", NaN)	Bank not assessed	`Bank evaluation missing`

Step 3: Missing Indicator Values

For each measured indicator (`indicator_cols`):

Condition	Explanation	Note Added
Value is NaN or blank AND `Comments` is empty, "n/a", "na", or "none"	Missing value without justification	`X missing without explanation` (`X` = indicator name)
Value is NaN BUT `Comments` has explanation	Only value is removed (set to `NaN`)	No error note added

Step 4: `"Image of site was submitted"` Check

Value	Action	Note
`"no"`, `"false"`, `"n"`, `""`, `"nan"`	Marked as image not submitted	`Site image not submitted`
`"yes"`, `"true"`, `"y"` or similar	Value standardized to `"Yes"`	`Image value standardized: '{val}' → 'Yes'`

Final Output Summary

File	Content
`cleaned_RIPARIAN.xlsx`	Only rows with empty `"RIPARIAN_ValidationNotes"`
`annotated_RIPARIAN.xlsx`	All rows with additional columns:

- `"RIPARIAN_ValidationNotes"`
- `"RIPARIAN_ChangeNotes"`
- Adjusted fields (e.g., standardized `"Image"` values, NaN entries) |