Safe Water
Optimization Tool

SWOT EO Version 1.5 Updates

Document Version 1.0
December 31, 2023
Michael De Santi, Dr. Usman T. Khan, Dr. Syed Imran Ali

Revision History

Date	Version	Notes
December 31, 2023	1.0	Original document

Introduction

The Safe Water Optimization Tool (SWOT) generates site-specific, evidence-based water chlorination targets that help ensure water safety up to the point-of-consumption in humanitarian response settings such as refugee camps. The SWOT unlocks operationally relevant insights from water quality monitoring data collected at the point-of-distribution (tapstands) and at the point-of-consumption (households after the typical duration of storage). The SWOT uses two modelling engines to generate this chlorination guidance: a probabilistic machine learning approach, the SWOT-ANN, and the Engineering Optimization (EO) tool, which uses process-based chlorine decay models. This white paper describes how the Version 1 EO was updated to produce the new Version 1.5 EO modelling engine of the SWOT.

To generate an FRC target, the SWOT EO receives paired and timestamped measurements of tapstand and household FRC uploaded by the user. The user specifies a storage duration (e.g., 15 hours) and a decay scenario (Optimal, Minimum or Maximum). The storage duration is the length of time that FRC needs to persist in household-stored drinking water (typically the maximum expected duration that water will be used). The decay scenario determines how conservative the FRC target will be. For example, if there are concerns about waterborne illnesses, substantial variations in expected storage duration, or poor environmental hygiene, the Maximum Decay scenario may be needed to provide additional protection.

Version 1 of the EO was developed in 2019 and uses the power decay model—an empirical reaction kinetics model—to simulate post-distribution decay. In version 1, the decay parameters were calibrated using a grid-search which tests 90,000 possible parameter combinations. “Optimal decay” parameters are identified that minimize the sum of squared errors (SSE, Equation 1). “Minimum Decay” and “Maximum Decay” parameters are then identified as the parameters that produce the least or most decay (respectively) within 5% of the optimal solutions’ error. The initial development of the EO engine and details of its functions are provided in Ali et al. (2021)[1].

$$SSE = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \qquad (1)$$

- Where y_i is the i^{th} observation and $\hat{y}_i$is the corresponding model prediction.

Since the official launch of the Version 1 SWOT, which used the Version 1 EO, in November 2021 the SWOT has been deployed in several humanitarian response settings around the world. While the EO has provided crucial evidence-based, site-specific FRC targets in these settings, several limitations of the Version 1 EO have been identified through these field deployments:

1. The grid search approach to parameter selection is slow and computationally expensive;
2. The power decay model occasionally fails to converge, particularly when there is a mismatch between elapsed time between paired FRC measurements and the FRC target storage duration;
3. The SWOT needs to be recalibrated every time an FRC target is needed for a different storage duration, which is cumbersome and time consuming;
4. There is a lack of interpretability of the modelling results.

These limitations motivated the design of new features in the Version 1.5 EO which was released in the Version 2 SWOT Launch in November, 2022.

[1] S. I. Ali, S. S. Ali, J. F. Fesselet, Evidence-based chlorination targets for household water safety in humanitarian settings: Recommendations from a multi-site study in refugee camps in South Sudan, Jordan, and Rwanda. *Water Research,* **189**, 116642 (2021).

Table 1: Summary of new features designed to overcome limitations of the Version 1 EO.

Limitation	***New Feature***
1. Slow grid search	**Fast ensemble modelling**
2. Poor convergence of the power decay model	**Multi-model training**
3. Retraining for new targets	**Improved model outputs**
4. Lack of interpretability of modelling performance	**Model and target confidence assessment**

New Features of the Version 1.5 EO

Fast Ensemble Modelling

The Version 1 EO used a grid search approach to define the optimal decay parameters (i.e., those that produce the lowest SSE) and minimum and maximum decay parameters (i.e., those that produce highest/lowest decay within 5% SSE of the optimal decay model). Grid search is an optimization approach that tests all possible combinations of a set of parameters. Since the power decay model used in the Version 1 EO is a two-parameter model, and 300 possible values are provided for each parameter, the grid search tests 90,000 parameter combinations. However, most of these parameter combinations produce poor performance (as the model parameters are very far from the optimal values) and all parameters that do not produce performance within 5% of the best model's error are discarded. Thus, the majority of the 90,000 tested parameter combinations are not used and do not produce good performance. This approach was taken because it is thorough and because it provides a good definition of the SSE across the parameter space. However, is computationally inefficient and time consuming. It also limits the ability of the SWOT to use other, more complex models, as each parameter added to the model increases the computational complexity by a factor of 300 (even a simple three-parameter model would produce 27 million parameter combinations).

To overcome these challenges, the Version 1.5 EO uses an ensemble modelling approach to calibrate the model parameters. In this approach, the EO calibrates the FRC decay model 100 times to obtain an ensemble of 100 unique sets of model parameters. Each calibration run begins with randomized initial parameters and is calibrated using only 75% of the overall dataset, using the remaining 25% as a holdout set for validating model performance set. In each run, a numerical solver iteratively updates model parameters to identify parameters that minimize SSE. Using a different subset of the data for calibration in each run and starting with different initial parameters promotes diversity in the final ensemble of model parameters. At the end of each run, the EO calculates the mean squared error (MSE) and coefficient of determination (R^2) for both the calibration and validation datasets (Equations 2 and 3).

$$MSE = \frac{SSE}{N} \tag{2}$$

- Where there are *N* observations

$$R^2 = 1 - \frac{SSE}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{3}$$

- Where $\bar{y}$ is the mean of the observations and y_i is as previously defined.

This ensemble approach is considerably faster than the grid search approach. There are a maximum of 600 iterations per run, so a maximum of 60,000 calculations are performed (in practice this is even lower since each run typically converges within 10 iterations). Using 75% of the dataset for calibration further reduces the computational demands and allows us to validate the performance of the calibration process on independent data (the remaining 25% of the data). Additionally, unlike in the Version 1 SWOT, no runs are spent on solutions far from the optimal region in the solution space. All computational resources are only spent on high-performing solutions.

These benefits are achieved without sacrificing the EO's ability to explore the error distribution across the parameter space. This distribution can be explored from the ensemble parameters. Since each run is calibrated on different subsets of the data, the ensemble contains a diverse set of final parameters, which the Version 1.5 EO uses to determine the "Minimum Decay" and

"Maximum Decay" parameters. The Version 1.5 EO selects the minimum/maximum decay parameters from the 75% of solutions with the lowest SSE. The "Maximum Decay" scenario parameters are those that produce the highest FRC target at the user-defined storage duration and the "Minimum Decay" parameters are those that produce the lowest FRC target at the storage duration.

A final benefit of the Version 1.5 EO's ensemble approach is that the maximum number of iterations is the same regardless of the number of model parameters. Thus, more complex models can be used without sacrificing computational time or resources.

Multiple Decay Models

Unlike the Version 1 EO which only uses the power decay model which can fail to converge to valid solutions, the Version 1.5 EO uses three different FRC decay models:

1. Power decay (as in the Version 1 EO);
2. First-order decay (Equation 5);
3. Parallel first-order decay (Equation 6).

The first-order decay model is simpler than the power decay model, with the decay order (*n)* fixed as one (first-order exponential). This equation contains only one free parameter: the decay rate (*k*). This simplicity can reduce the performance of the first-order model as it cannot fully reflect the behaviour at a given site; however, the first-order model does not have the same overfitting and convergence risks as more complex models.

$$C = C_0 + e^{-kt} \tag{5}$$

The parallel first-order decay model is an extension of the first-order model. It assumes that the underlying chlorine decay is represented by two first-order reactions occurring in parallel: a slow reaction with a decay rate of k_1 and a fast reaction with a decay rate of k_2. These two reactions are assumed to occur in parallel and not conflict with each other, so the overall decay is split between the two reactions based on *w.* The proportion of chlorine decaying at the slow rate is *w*, and *(1-w)* is the proportion of chlorine decaying at the fast rate. The advantage of the parallel first-order model is that it provides additional complexity not present in the first-order model while avoiding the convergence issues that can occur in the power decay model where *n* is a free parameter. However, this model can still fail to converge, particularly if k_1 drops to 0 as this artificially produces a limited first-order decay reaction where the slow decay portion (*w*) does not proceed.

$$C = \boldsymbol{w} * C_0 * e^{-\boldsymbol{k_1}t} + (1 - \boldsymbol{w}) * C_0 * e^{-\boldsymbol{k_2}t} \tag{6}$$

Convergence Checks and Model Selection

The Version 1.5 EO implements convergence checks to identify and remove solutions that either overfit the training data or fail to converge. The Version 1.5 EO's calibration approach produces ensembles of 100 sets of model parameters for each decay model. The first convergence check reviews the parameters of the individual runs and removes individual power decay and parallel first-order runs where *n* (power decay) or k_1 (parallel first-order) are less than 10^{-5} (a small number substitute for 0). In the power decay model this produces non-exponential decay, and in the parallel first order model this produces a limited first order decay model. After removing these runs, the convergence checking process evaluates the following criteria for both the full ensembles of calibrated parallel first-order and power decay models. We deem that a given model has failed to converge or is overfitted if any of the following checks are true:

- There are no remaining parameter sets after removing individual runs;
- The variance of any model parameter is greater than 0.1 (indicates poor convergence);
- The difference between the maximum and minimum decay scenario FRC targets at the user-specified storage duration is greater than 1 mg/L (indicates poor convergence);
- The FRC target for the user-selected decay scenario is below 0.3 mg/L or above 100 mg/L (indicates poor convergence);
- The FRC target for the user-selected scenario returns NaN (not a number, which typically occurs when an imaginary number is encountered, indicating poor convergence);
- The R^2 or MSE is more than 50% worse in validation than in calibration for either the optimal decay model or for the 75th percentile model (indicates overfitting to the calibration data).

After applying these checks to assess convergence and overfitting, the Version 1.5 EO selects the model that passes these convergence checks with the lowest MSE and uses the selected model to generate the SWOT FRC target. If no model passes the convergence check, the Version 1.5 EO reverts to using the first-order model as a fail-safe approach, since this is the least complex model and will always converge to a solution under normal circumstances.

Improved Model Outputs

In addition to improving the performance of the EO, the Version 1.5 EO also provides an updated set of outputs to better communicate FRC target recommendations and calibration results to users.

FRC Targets for Multiple Storage Durations

There are many cases where a user may need or want multiple FRC targets for the same dataset at different storage durations. For instance, they may have received a SWOT target that is too high to be acceptable to water-users at their site; they may have multiple storage duration peaks; or they may be planning a change in water supply approaches that will change household storage behaviours. For a given set of model parameters, an FRC decay equation can be used to determine the required FRC target for any storage duration without retraining. However, in the Version 1 EO, users must rerun the entire SWOT Analysis function to obtain an FRC target for a new storage duration. In the Version 1.5 EO, we have added additional model outputs so that users can directly identify the required target for different storage durations. Figure 1 shows the required target over time for a user-selected decay scenario. This output is provided on the *Results* page of the Version 2 SWOT web app (www.safeh2o.app). This figure shows users what FRC target they would need for different storage durations, meaning that there is no need to rerun the SWOT. We also provide Figure 1 in tabular form and include the selected model and decay parameters in the final report so that users can calculate an FRC target for any storage duration that they want using the appropriate decay equation.

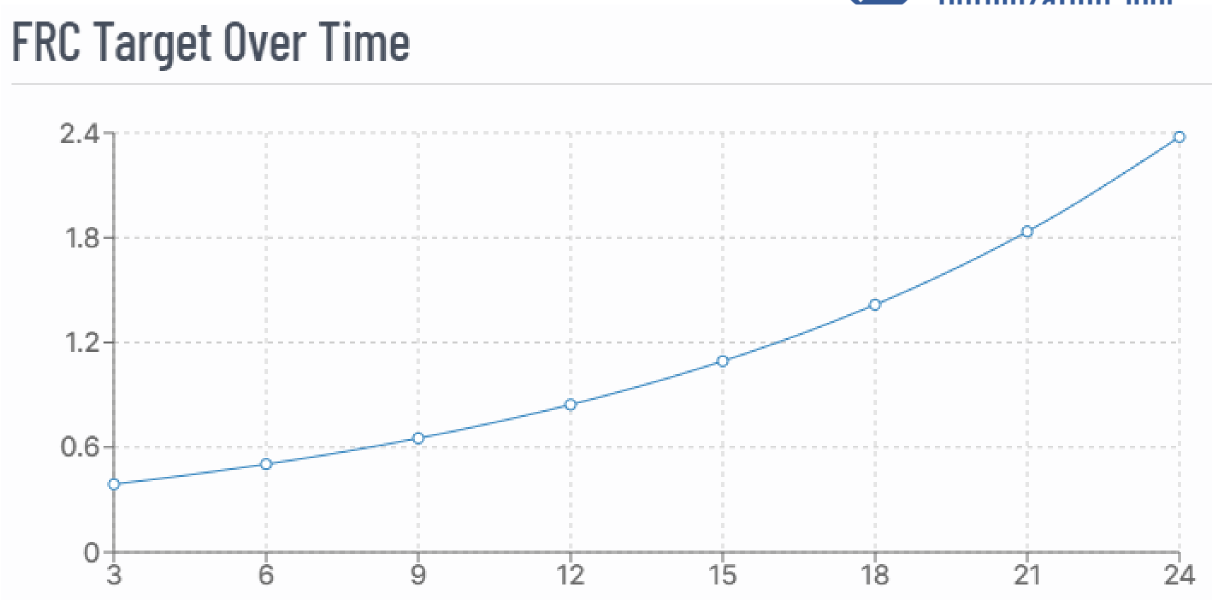

Figure 1: This figure, shown on the Version 2 SWOT v2 Results page for an Analysis run, and in the detailed report for the Version 1.5 EO, shows the required FRC concentration to provide FRC protection for different storage durations, so that the model does not need to be rerun.

Duration of FRC Protection

We also show the projected decay of the existing FRC target over storage durations longer than the user-specified storage duration. This allows users to identify how long FRC will persist beyond the user-specified storage duration if the SWOT target is met at the tapstand. Note that the FRC starts at 0.3 mg/L because the SWOT always optimizes targets for 0.3 mg/L FRC at the household to provide a factor of safety above the required 0.2 mg/L due to variability in post-distribution FRC decay.

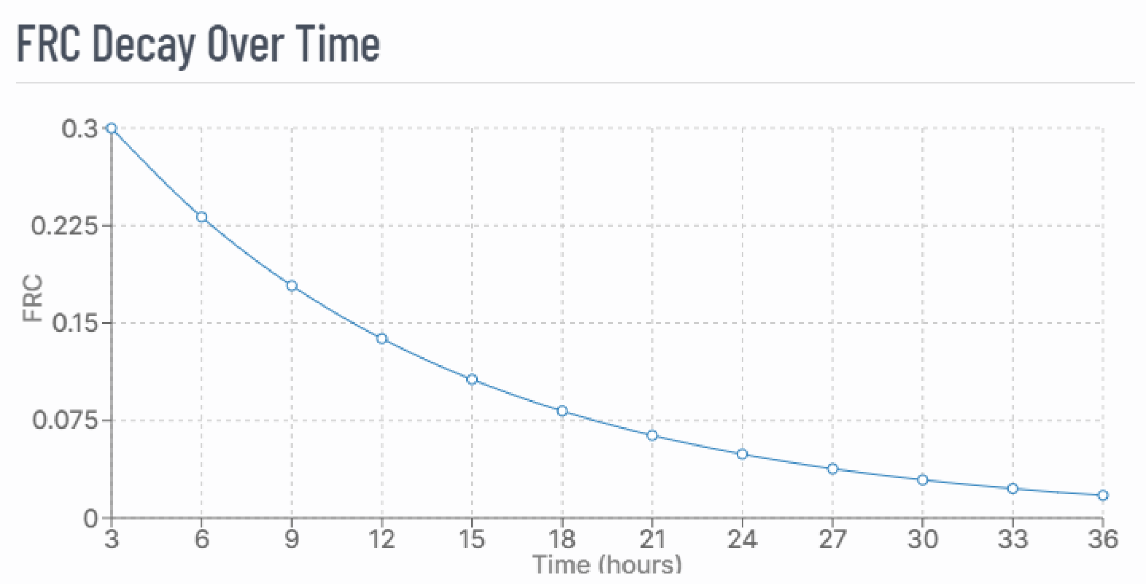

Figure 2: New SWOT EO output shows projected household FRC concentration for storage durations longer than the user-specified storage duration when the SWOT EO target is achieved at the tapstand. The decay curve starts at 0.3 mg/L at the user-specified storage duration because SWOT targets are calibrated with a factor of safety to account for

variability in FRC decay. Here we see that even with a specified storage duration of 12 hours, the average water container will have safe water up to 18 hours.

Model and Target Confidence Assessment

The features described so far were designed to solve functionality problems identified in the Version 1 EO. However, the Version 1 EO also lacked a clear mechanism for communicating to users when the EO was not performing well or not functioning as intended. In response to this challenge, the Version 1.5 EO includes a confidence assessment to communicate issues to users using a confidence rating of "High/Medium/Low". The bulleted lists at the bottom of each of the following sub-sections provide the possible confidence ratings for each item in the confidence assessment as well as the user message.

Number of Samples

The first item in the confidence assessment is the number of observations used for modelling. We recommend 100-150 samples to run the EO. Thus, the following criteria are used to rate this item in the confidence assessment:

- Low Confidence: Fewer than 100 observations
- Medium Confidence: 100-150 observations
- High Confidence: More than 150 observations

Stability of Household and Tapstand FRC

The second item in the confidence assessment is the stability of the household and tapstand FRC data. The Version 1.5 EO evaluates this by randomly sampling 10% of the dataset 100 times and evaluating the similarity of the variance and distribution of the sampled 10% from the remaining 90% of the data. This assessment provides a further insight on whether enough samples have been collected. If the difference in variance and/or distribution between the sampled 10% and remaining 90% is significant, this indicates that the dataset may not be adequately capturing the decay behaviour at the site. The EO uses a Levene test to evaluate the similarity of variances and a two-sample Kolmogorov Smirnov test to measure the similarity of the distributions. The EO uses these statistics to evaluate differences between the two groups for both household FRC and tapstand FRC there are a total of four tests:

1. Levene test for tapstand FRC;
2. Levene test for household FRC;
3. Kolmogorov-Smirnov for the tapstand FRC;
4. Kolmogorov-Smirnov for the household FRC.

Both the Levene and Kolmogorov-Smirnov tests are hypothesis tests with a null hypothesis that there is no difference between the two datasets. Thus, more similar distributions/variances will have higher p-values. For the confidence assessment, we take the minimum p-value across all 100 iterations of all four tests and use the following rating system:

- Low Confidence: Variance and distribution of household and tapstand FRC are significantly different at a p-value level of 0.05;
- Medium Confidence: Variance and distribution of household and tapstand FRC are not significantly different at a p-value 0.05 but are significantly different at a p-value of 0.10;
- High Confidence: Variance and distribution of household and tapstand FRC are not significantly different at a p-value level of 0.10.

Alignment Between Elapsed Time in Collected Data and User-Specified Storage Duration

The third item in the confidence assessment is the alignment between the elapsed time between tapstand and household measurements in the paired samples and the user-specified storage duration target. It is important that the user-specified storage target and observed storage durations align because the calibrated models will be most accurate wherever there are the most observations (since overall SSE can be minimized by having very accurate predictions where there is a large amount of data even with very poor predictions elsewhere). The EO rates the confidence based on the proportion of observations with elapsed time within +/- 1.5 hours of the user-specified storage duration. The thresholds for number of observations to achieve high, medium, or low confidence were determined based on experiences using data sets provided for SWOT research. Note that this item in the confidence assessment should be considered in the context of the next check (uniformity of distributions). If the storage durations are uniformly distributed across time periods, then the alignment between elapsed times in the data and the user specified storage duration are less critical.

- Low Confidence: Fewer than 50% of observations have sampling durations within +/- 1.5 hours of the user-specified target storage duration OR No observations have sampling durations within +/- 1.5 hours of the target storage duration.
- Note: the second condition is a subset of the first, however the message provided to users varies depending on which condition is met.
- Medium Confidence: Proportion of observations within +/- 1.5 hours of storage target is higher than the median proportion of observations in all groups.
- High Confidence: Proportion of observations within +/- 1.5 hours of the storage target are higher than the upper quartile proportion of observations in all groups.

Uniformity of Storage Durations

The fourth item in the confidence assessment also reviews the distribution of storage durations in the observations, but this time for uniformity. As noted above, the more observations there are within any storage duration group, the more that group contributes to SSE, which can lead to models becoming overly specific to a single storage duration group. Thus, a uniform distribution of elapsed times in the paired data allows the model to generalize FRC targets for a broader range of storage durations. The EO measures the uniformity of the distribution of storage durations using a two-sample Kolmogorov-Smirnov test between the actual distribution of observations and a uniform distribution. This test only evaluates storage periods less than or equal to the +/-1.5-hour range around the user-specified storage duration target since longer-term storage may be infeasible at a site and including them in the evaluation would incorrectly lower the confidence assessment.

- Low Confidence: p-value lower than 0.05. Minimal uniformity of distribution of samples up to target storage duration
- Medium confidence: p-value between 0.5 and 0.05. Mostly uniform distribution of samples up to target storage duration
- High Confidence: p-value above 0.5. Uniform distribution of samples up to target storage duration

FRC Decay Magnitude

The fifth item assesses the magnitude of FRC decay at the site. Using the selected decay scenario, the EO calculates the target FRC at 24 hours. If the FRC target is above 2 mg/L, users are warned that the decay is higher than expected. This can occur normally (due to poor environmental hygiene or hot temperatures), but it can also occur if only short-duration samples are collected because FRC decay is fastest immediately after water is collected and then slows down over time. This confidence assessment is a binary measure (the target is either above or below 2 mg/L) so there are only “High” and “Low” confidence ratings. The “Low” confidence rating message also includes a caveat that this is only concerning if the majority of samples have storage durations less than 6 hours because in these cases, the high magnitude of FRC decay may be biased due to FRC decay typically being faster earlier in storage.

- Low Confidence: Target at 24 hours is >2 mg/L. FRC decay parameters are higher than normal. This may lead to higher FRC than acceptable FRC targets at long storage durations. This can occur normally in some sites, or it may be due to a large number of short-duration samples. If the average storage duration is shorter than 6 hours, please collect additional samples at longer storage durations to see if this condition persists.
- High Confidence: Target at 24 hours < 2 mg/L. FRC decay parameters are reasonable.

Model Performance

The sixth and final item in the confidence assessment is the model performance. The EO evaluates performance using the validation R^2. The validation R^2 is used because this indicates of generalizable performance (the expected performance on new data). We selected R^2 both as it is well known and because the formulation we use, which is equivalent to the Nash Sutcliffe Efficiency (NSE), has an interpretable meaning: where an R^2 or NSE of 0 means the model predicts as well as predicting the mean, above 0 means the model is better than predicting just the mean, and below 0 means that it predicts worse than the mean.

- Low Confidence: R2 <0. Model performance is poor.
- Medium Confidence: R2 between 0 and 0.8. Model performance is acceptable.
- High Confidence: R2 above 0.8. Model performance is good.

Next Steps

The Version 1.5 EO provides several key improvements over the Version 1 EO. However, this is an intermediate step towards further evolutions of the SWOT EO. Future potential changes are:

- Updating the fast ensemble calibration approach to be probabilistic instead of selecting a single model to better quantify the variability in FRC decay;
- Use advances documented in literature to modify decay rates for water temperature based on Arrhenius' equation as a step toward making the model sensitive to water quality parameters;
- Incorporate more advanced solvers, such as genetic algorithms and particle swarm optimization, to obtain better performance and identify global solutions;
- Add additional aspects of the confidence assessment based on use cases and user feedback; and
- Implement dynamic outputs so users can query model results.