Methods for Automated Generation of Field Scenarios and Postprocessing of Pesticide Water Calculator Output

U.S. Environmental Protection Agency
Office of Pesticide Programs

12/31/2019

Table of Contents

1. Introduction

As a part of the requirements for pesticide registration and registration review, the U.S. Environmental Protection Agency, Office of Pesticide Programs (OPP) conducts aquatic exposure assessments to determine whether pesticides that are applied according to label directions may result in concentrations in water that may adversely impact human health or aquatic organisms. OPP estimates pesticide concentrations in water with models that simulate soil, weather, hydrology, and management/use conditions that are expected to influence the potential for pesticides to move into water.

These models – the Pesticide in Water Calculator (PWC) (Young, 2019), the Pesticide Root Zone Model (PRZM5) (Young and Fry, 2016), which is a component of the PWC, and the Spatial Aquatic Model (SAM) (USEPA OPP, 2015) – use scenarios to represent field, watershed, and waterbody properties that are important in pesticide fate and transport. Field scenarios refer to the field and watershed (environmental) inputs used in the models. USEPA OPP (2019) describes field scenario inputs used in the models. This document describes the code used to generate field scenarios and to compile the inputs for the aquatic models.

2. Field Scenario Generation

Scenarios generation was automated through the execution of a series of scripts developed in the Python programming language (Figure 1). These scripts are managed and hosted in the GitHub version control system, in a public repository within EPA's GitHub account (https://github.com/usepa/opp-efed). The code was written as simply and cleanly as possible for clarity and brief execution times, and the GitHub repository and the code itself contain markup and documentation to explain the operations in the scripts.

There are two steps in the automated generation of scenarios:

1. Spatial overlay of GIS layers to generate spatial index
2. Tabular join of input datasets to spatial index and collation of scenario groups

A third and final step, Processing of PWC output files and scenario selection by estimated exposure, is explained in Section 3.

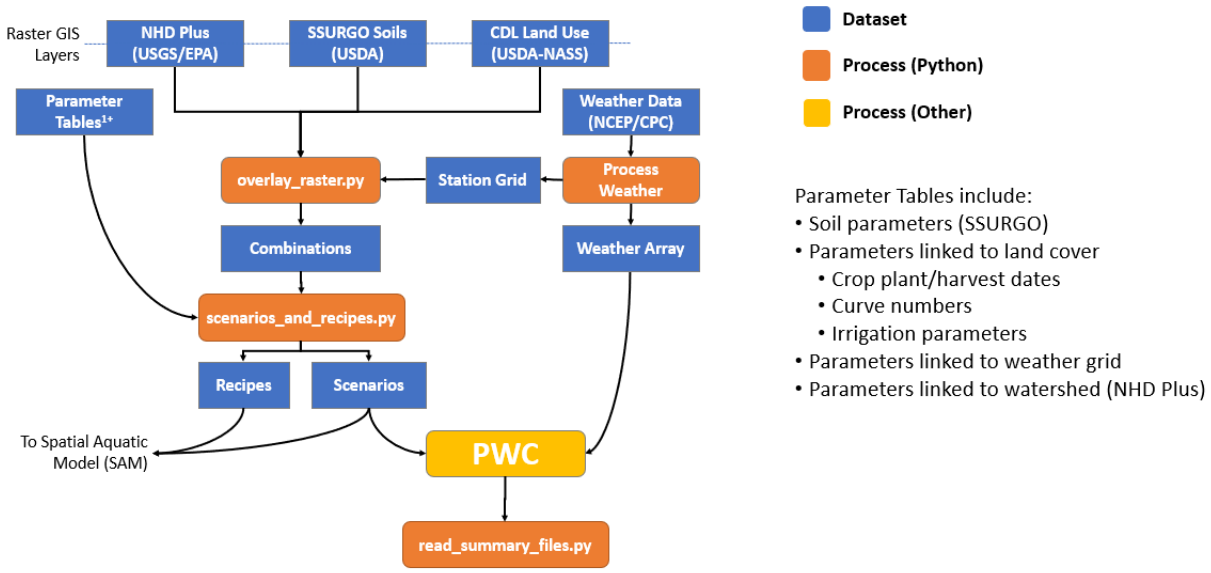

Figure 1. Aquatic model input generation flowchart

2.1. Step 1: Spatial Data Overlay

The script **overlay_raster.py** generates a spatial index from the overlay of 3 raster (gridded) GIS datasets: (1) SSURGO map unit ID (USDA NRCS SSS, 2018), (2) Cropland Data Layer (CDL) crop ID (USDA NASS, 2014-2018), and (3) weather grid ID (Fry et al., 2016). All spatial data except SSURGO had a 30 m cell size and a national (contiguous 48 states) extent. The raster SSURGO datasets were disseminated at the state extent and a 10 m cell size; thus, the script performs a spatial merge and resampling on these layers with ArcGIS prior to the overlay. The script utilizes the ArcPy module (ArcGIS Pro 2.3) to perform a Combine operation on the input raster datasets. The result is a raster dataset, and corresponding attribute table, containing three identifiers (referred to hereafter as index fields) for each pixel in the contiguous United States: soil map unit ID (*mukey*), CDL crop ID (*cdl*), and weather grid ID (*weather_grid*), and an area field containing the spatial extent of each unique combination. The attribute table, named the *combinations* table in the code, constitutes the spatial index from which the scenarios are built.

2.2. Step 2: Population of Scenario Parameters

The script **scenarios_and_recipes.py** generates scenarios from the spatial index. The script first aggregates combinations for all years by combining rows with common index field values. This aggregation also combines the values of the *area* field for grouped combinations. To account for double-cropped classes in the CDL (e.g., CDL class 26 – Winter Wheat/Soybeans), the script duplicates combinations with a double-cropped CDL crop ID and assigns each of the duplicated combinations a *cdl_alias* using the CDL crop ID for each of the constituent crops. An example of the use of this alias is where a single combination with *cdl* = 26 would be duplicated, with one of the duplicates assigned *cdl_alias* of 24 (winter wheat) and the other a *cdl_alias* of 5 (soybeans). Both rows would retain the

original *cdl* value of 26. This *cdl_alias* is used as an indexing field for most crop-linked parameters. Another index field, *state*, is created from a one-to-one relationship with soil map unit ID.

At this point, the combinations table has 5 index fields *cdl*, *weather_grid*, *mukey*, *cdl_alias*, and *state*. The script creates scenarios by joining several tabular datasets indexed by one or more index fields to the combinations table. These tables and the fields contained within are listed here (index fields in bold, full parameter descriptions in *Estimating Field and Watershed Parameters Used in USEPA's Office of Pesticide Programs Aquatic Exposure Models* (USEPA OPP, 2019):

- **cdl_params.csv**
  - ***cdl, cdl_alias****, gen_class, label_group, crop_intercept, max_cover, amxdr, usle_c_fal, usle_c_cov, cultivated*
- **crop_dates.csv**
  - ***cdl, cdl_alias, state****, bloom_begin, bloom_end, plant_begin, plant_begin_active, plant_end_active, plant_end, harvest_begin, harvest_begin_active, harvest_end_active, harvest_end*
- **curve_numbers.csv** (indexed to *gen_class* parameter from crop_params.csv)
  - ***gen_class****, cn_fal_A, cn_fal_B, cn_fal_C, cn_fal_D, cn_cov_A, cn_cov_B, cn_cov_C, cn_cov_D*
- **irrigation.csv**
  - ***cdl, state****, irrigation_flag, irrigation_type, irrigation_rate, depletion_allowed, leaching_fraction*
- **met_params.csv**
  - ***weather_grid****, anetd, ireg*

The script takes parameters linked to soil directly from the SSURGO dataset. There were three tables within the state-level SSURGO data and one national-level table containing parameters used for scenario generation (index fields in bold, SSURGO field headers in parentheses, field descriptions in (USEPA OPP, 2019)):

- **muaggatt** – Properties indexed to soil map unit
  - ***mukey****, hydro_group_dominant (hydgrpdcd)*
- **component** – Properties indexed to soil component
  - ***mukey, cokey****, component_pct (comppct_r), hydro_group (hydgrp), major_component (majcompflag), slope (slope_r), slope_length (slopelenusle_r)*
- **chorizon** – Properties indexed to soil horizon
  - ***cokey****, bd (dbthirdbar_r), fc (wthirdbar_r), wp (wfifteenbar_r), orgC (om_r), sand (sandtotal_r), clay (claytotal_r), horizon_top (hzdept_r), horizon_bottom (hzdepb_r), horizon_letter (desgmaster), pH (ph1to1h2o_r), usle_k (kwfact)*
- **valu1** – National-level data indexed to soil map unit
  - ***mukey****, root_zone_max (rootznemc)*

In SSURGO, a map unit describes soils and other components that have unique properties, interpretations, and productivity. A single map unit may contain multiple components, one or more of which is designated as the 'major component' for the map unit. For scenario generation, the script chooses the major component comprising the largest area of the map unit (*component_pct*) to

represent the entire map unit. Soil horizon data are linked to the component table. Quality control checks are applied to the soils data to identify missing or out-of-range values in soil horizons data. If an invalid or missing value is encountered in a soil horizon (e.g, *horizon_bottom* = 0), that horizon and all horizons below it in the map unit are deemed invalid, and the soil profile is truncated at the deepest valid horizon. The adjusted soil profile depth is applied as a maximum value to the *root_zone_max* parameter.

Additional parameters are derived from SSURGO parameters (field descriptions in (USEPA OPP, 2019)): *thickness*, *root_depth*, *evaporation_depth*, *n_horizons*, *usle_k*, *usle_ls*, and *usle_p*.

Additional parameters are also computed from combinations of other parameters based on the results of the join (field descriptions in (USEPA OPP, 2019)):

- *plant_date*, *emergence_date*, *maxcover_date*, and *harvest_date* are derived from other crop date fields (e.g., *plant_date_begin, plant_date_end*)
- *cn_cov* and *cn_fal* are selected from curve number fields (e.g., *cn_fal_A*, *cn_cov_D*) based upon the combination of CDL class (*cdl_alias*) and hydrologic soil group (*hydro_group*)

Once the scenarios are fully populated with all required parameters, the script performs a quality control check to identify and remove scenarios with invalid or missing values. The expected values for the inputs were based primarily on documentation for the source data as described in USEPA OPP (2019). The table **fields_and_qc.csv** is provided as an input to the script and containing the name of each field, along with a set of QC parameters: *range_min/max*, *range_flag*, and *blank_flag*. These parameters are used to generate flags based on whether the values in each field are missing, fall out of range, or fall out of a more conservative 'general' range. The value of the flag represents the severity of the error. The *flag* parameters may have a value of 0, 1, or 2, where 0 represents no error, 1 represents an unusual value that does not invalidate the scenario, and 2 represents an error which renders the scenario invalid. Scenarios where any field has a flag of 2 are removed from the scenarios table. An output file (**r[*region*]_qc.csv**) is generated that provides the flags for each scenario and field, and another (r[region]_report.csv) is generated that provides the number of scenarios marked for removal by a flag of 2 or higher for each field.

The scenarios table with all valid scenarios is written to file (**r[*region*]_parent.csv**). This table is then collated and sampled to produce a random sample of scenarios for each region and crop or crop group.

3. Step 3: PWC Output Postprocessing

The final scenarios table for each region and crop (r[region]_parent.csv) was batch processed in PWC to simulate estimated drinking water concentrations (EDWCs). PWC generates a summary file (**BatchOutputVVWM.txt**) containing modeled pesticide concentrations at different exposure durations (e.g., acute, chronic and cancer), indexed by scenario ID. A post-processing script (**read_summary_files.py**) was used to read this output file, rank the scenarios by concentration and assign percentiles, and return a selection of scenarios corresponding to a selection percentile. This script also generates plots of the distribution of modeled concentrations.

The script first reads the PWC output file, then performs a tabular join with the original scenarios file, retaining scenario ID, area, and hydrologic soil group fields. This table is then used to calculate percentiles for each scenario. To produce the rankings, the script sorts the estimated acute, chronic, and cancer EDWCs from the field scenarios for each HUC2-region/K_{oc} combination simulated by PWC from highest to lowest and produces percentiles for the ranked field scenarios according to the following equation (NIST 2019, Wikipedia 2019)

$$p_n = \frac{\sum_{i=1}^{n} A_i - \frac{A_n}{2}}{\sum_{i=1}^{N} A_i}$$

Where p_n is the percentile ranking at a given scenario *n*, $\sum_{i=1}^{n} A_i - \frac{A_n}{2}$ is the cumulative area of one half of scenario *n* and all scenarios ranked below scenario *n*, and $\sum_{i=1}^{N} A_i$ is the total area of all scenarios. The interpretation of the 90th percentile would be that 90% of the cropped area, as represented by all field scenarios ranked below the 90th percentile, will result in lower estimated concentrations.

The percentile values are appended to the combined scenario/PWC output table. After assignment of percentile values, the complete table is written to file (**[*run id*]_summary.csv**).

The scripts performs scenario selection for each selection percentile, which is the target percentile for the representative scenario. The scenarios with percentile values closest to the selection percentile are selected and written to file (**[*run id*]_selection.csv**). The script generates plots written to files (**plot_[*duration*].png**) for the distributions of concentration and percentiles for each exposure duration (i.e., acute, chronic, cancer).

4. References

Fry, M.M., G. Rothman, D.F. Young, and N. Thurman. 2016. Daily gridded weather for exposure modeling. Environmental Modelling & Software, 82, 167-173, doi.org/10.1016/j.envsoft.2016.04.008

NIST/SEMATECH e-Handbook of Statistical Methods, 2019. https://www.itl.nist.gov/div898/handbook/prc/section2/prc262.htm

USDA National Agricultural Statistics Service Cropland Data Layer. 2014-2018. Published crop-specific data layer [Online]. Available at https://nassgeodata.gmu.edu/CropScape/ (accessed Feb 2019). USDA-NASS, Washington, DC.

USDA Natural Resources Conservation Service Soil Survey Staff (USDA NRCS SSS). 2018. Gridded Soil Survey Geographic (gSSURGO) Database. United States Department of Agriculture, Natural Resources Conservation Service. Available online at http://datagateway.nrcs.usda.gov/. October 11-12, 2018 (FY2018 official release).

U.S. Environmental Protection Agency Office of Pesticide Programs (USEPA OPP). 2015. Development of a Spatial Aquatic Model (SAM) for Pesticide Risk Assessments. Presented to the FIFRA Scientific Advisory Panel, September 15-17, 2015. Available in the public e-docket, Docket No. EPA-HQ-OPP-2015-0424, accessible through the docket portal: http://wv.rw.regulations.gov.

U.S. Environmental Protection Agency Office of Pesticide Programs (USEPA OPP). 2019. Estimating Field and Watershed Parameters Used in USEPA's Office of Pesticide Programs Aquatic Exposure Models –

The Pesticide Water Calculator (PWC)/Pesticide Root Zone Model (PRZM) and Spatial Aquatic Model (SAM). Draft.

Young, D.F. 2019. The USEPA Model for Estimating Pesticides in Surface Water, in *Pesticides in Surface Water: Monitoring, Modeling, Risk Assessment, and Management*. American Chemical Society, editors Goh, Kean, and Young, American Chemical Society, Washington DC.

Young, D.F. and Fry, M.M. 2016. PRZM5: A Model for Predicting Pesticide in Runoff, Erosion, and Leachate, Revision A. USEPA/OPP 734S16001, U.S. Environmental Protection Agency, Washington, DC. Available for download with the Pesticide in Water Calculator from the USEPA OPP Water Models web site at https://www.epa.gov/pesticide-science-and-assessing-pesticide-risks/models-pesticide-risk-assessment#aquatic

Wikipedia contributors. Percentile. Wikipedia, The Free Encyclopedia. 2019. https://en.wikipedia.org/wiki/Percentile#The_weighted_percentile_method