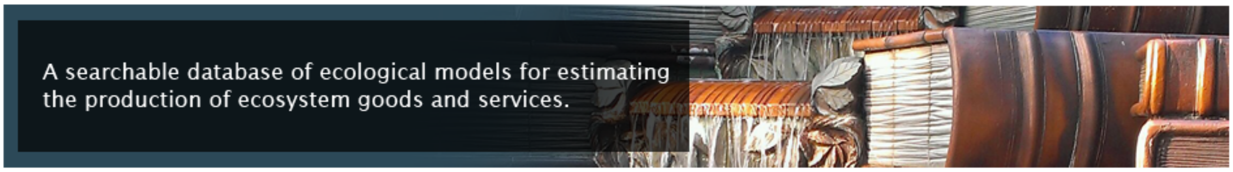
EcoService Models Library (ESML)



US-EPA

Ecological Model (EM) Submission Guidelines for ESML

Data and metadata guidelines for submission

Introduction

Thank you for your interest in proposing an Ecological Model (EM) for inclusion in the the EcoService Models Library (ESML)! This document was created as a companion to the EM Submission Template. Please follow the 3 steps outlined below and the ESML Data Submission Group will consider your EM for publication in ESML.

ESML publishes EM descriptions through the US EPA's website, and as such, metadata and documentation of data are required for data publishing. Associated metadata must be validated when data are submitted to ESML. This step is important because many fields populated in the metadata will be used in ESML model descriptions and published to EPA's ESML website. The ESML Data Submission Group will work with the data submitter to verify data and metadata. Once data and metadata are verified and accepted, EMs will be published in a development environment for testing and placement in ESML. At this point, the data submitter may be able to review EM results depending on access to the EPA network. Input from data submitters will be sought, but final decisions about Variable Relationship Diagrams as well as EM and variable naming conventions will be completed by the ESML Data Submission Group. Once an EM's development has been finalized and approved, the EM will be deployed to EPA's ESML website for public access and will be searchable based on content in the metadata record.

Guidelines for Submitting Data for Inclusion in ESML

In an effort to complete timely updates to ESML and to follow the guidelines required for publishing data to the EPA's ESML website, we request the following initial steps be used for submitting processed data:

Step 1: Describe the EM source document

Step 2: Identify EM variables and units

Step 3: Draft a Variable Relationship Diagram (optional)

Please send your completed EM Submission Template to **Newcomer-Johnson.Tammy@EPA.gov with the subject line: "EM Nomination."**

If the ESML Data Submission Group determines that the submission would be an appropriate and beneficial addition to ESML, you will be invited to submit a full EM and provided with further instructions for describing the EM(s) and the variables.

Specific instructions for submitting data to ESML continue below.

Step 1: Describe the EM Source Document

A. If the source document is already in ESML, list the document ID number in the EM Submission Template “Step 1” spreadsheet.

B. If the source document is new to ESML, leave the document ID number blank and fill out the rest of the columns on the EM Submission Template “Step 1” spreadsheet.

Source documents describing an EM and/or its application may include journal articles, descriptive materials such as user manuals from a modeling website, or project reports (including EPA reports). In some cases, multiple source documents (such as a journal article and a user’s manual) provide relevant information for a single model.

To qualify as sources for EMs entered to ESML, sources should have been peer reviewed or will be peer reviewed. For models made available on websites, while the website itself (or associated user manual) may not have undergone a peer review, there should be clear evidence of related publications (such as journal articles using the method or a closely related method) that have undergone peer review.

Step 2: Identify EM Variables

List the variables in EM Submission Template “Step 2” spreadsheet and fill out the applicable columns. The follow sections provide additional information on variable naming conventions, variable types, units, temporal characteristics, and spatial characteristics. The temporal and spatial characteristics are optional but encouraged.

Variables that are used by the EM must be identified so that they can be properly named, categorized and entered into ESML. Variables include most terms or quantities used in the model's mathematical or other logical formulations (exceptions are described below). In general, these include all terms required as inputs (called "predictors" in ESML) and all terms generated and used as outputs (called "responses").

Variables that are instructions to the model itself should normally be omitted. For example, a "switch" variable that determines whether a particular model function is turned on or off or variables that identify the time step or spatial pixel size should not be named and included as model variables (although spatial and temporal resolution will be defined by model or variable descriptors).

Certain types of model terms are considered trivial and are omitted. Mathematical constants (such as e, the base of natural logarithms and pi, the ratio of circumference to diameter) can be ignored; unit conversion factors (such as the number of centimeters per meter or number of days in a month), though important to computation, are considered informationally trivial and are ignored.

Constants, parameters, or coefficients used in the model that are not trivial by the above definition should be identified as variables. When they are numerous, however, and especially when their values are supplied by the model and they require little attention from the potential user, they may be named as a group rather than individually. For example, a single "variable" called "regression coefficients on evapotranspiration" could substitute for the individual naming of multiple coefficients in a model equation that estimates gross ecosystem production from ET.

Variables should be named to reflect data type, not data source. For example, if a model uses land use/land-cover class data from the National Land Cover Dataset, the variable should be called "Land use/land-cover class" and not "National Land Cover Dataset." (Data source for each variable is described in the variable's descriptors, not in its name.)

Intermediate terms wholly determined by the model equations and not regarded as outputs may be included as "computed intermediates," if in the enterer's judgment they have objective meaning (i.e., outside the immediate computational context) and may help clarify modeling context. For example, a water quality model may use an equation to estimate reach-specific stream discharge from mapped catchment data, and then use discharge in the computation of water quality. In this context, discharge is a computed intermediate; it does not have to be obtained prior to the use of the model, nor would it be considered a response variable, but it is objectively meaningful.

Variable Naming Conventions in ESML

When determining variable names, begin with the author's terminology but revise as needed to create a name that is understandable to a user who will see it in a dropdown menu, without surrounding context. Do this by applying the following general rules, in combination with common sense:

- Identify both a measure (e.g., "population density") and the thing to which the measure applies ("wolf")
- Note that "measure" does not mean unit, but perhaps number, mass, length, concentration etc.
- However, one or both measure or thing can be omitted if they are obvious simply going by the variable name. For example, Subsurface runoff can be assumed to be water, and while the measure could be either depth or volume -- technically different measures -- these both equate to volume measures which the units will clarify, so difference is trivial
- Note that "Soil emission rate" lacks the thing because we don't know what is emitted (it is not "soil")
- When adding thing, append in parentheses, e.g., "Soil emission rate (ammonia)"; "Age class composition (fish population)"
- If adding measure, do not use parentheses; e.g., rather than "impervious cover" use "area of impervious cover" or "percent impervious cover"
- For categorical variables (where measure does not apply) it may help to add "type" or "class;" e.g., instead of "land use," say "land use class" to distinguish from, e.g., an area
- For dummy variables (e.g., switches as in categorical regression) include "y/n" to indicate [e.g., "Pasture land use (y/n)"]
- Insert term like "index" where needed, e.g., "Quality of bathing water" is unclear and possibly should be "Quality index of bathing water," or perhaps "Bathing water quality index"
- If applications for multiple species are input as individual EMs, include species name in names of response variables so these EMs can be easily distinguished (not necessary for predictor variables)
- Variables should be named to reflect data type, not data source. For example, if a model uses land use/land-cover class data from the National Land Cover Dataset, the variable should be called "Land use/land-cover class" and not "National Land Cover Dataset." (Data source for each variable is described in the variable's descriptors, not in its name.)
- Spell out or replace any acronyms that would not be clear when viewed outside the context

Variable Types in ESML

ESML uses the term predictor to refer to independent variables (model inputs) and response for dependent variables (model outputs). Some models also have intermediary variables that are calculated from predictor variables in order to produce response variables.

This category is prefilled with the following options:

- **Predictor:Time- or space-varying**: Variable that tends to change with space and time and is causally related to (i.e., "drives") the response variable. Typically, must be measured or obtained from a dataset that is spatially and/or temporally specific. (Excludes any computed intermediate variables.)
- **Predictor:Constant or parameter:** This can be a factor that is relatively constant over space and time and modifies a driving variable, e.g. as a scalar multiplier or exponent applied to the driving variable. Sometimes obtainable by lookup (i.e., from measurement or estimation from a similar setting), by statistical estimation or by calibration. This could also be a categorical variable (e.g. land use class); though a categorical constant's value may vary over space or time, it remains constant for each class of this categorical variable. For example, a C sequestration rate, though varying across the landscape, for the purposes of the model, may be held constant (determined by) each designated land use class.
- **Intermediate (Computed):** Variable that is mathematically computed within the model yet not considered a response variable. If model computation takes place in stages, an Intermediate is a variable computed at an early stage and then used in later stages. For example, a water quality model may use an equation to estimate reach-specific stream discharge from mapped catchment data, and then use discharge in the computation of water quality. In this context, discharge is a computed intermediate; it does not have to be obtained prior to the use of the model, nor would it be considered a response variable. Computed intermediates are reported only if they are independently meaningful variables – i.e., they hold some relevance beyond the immediate context of the EM.
- **Response:Computed:** Response variable whose values are estimated by mathematical computation, using the model.
- **Response:Observed:** Response variable whose values are obtained by observation, not by computation using the model. The primary situations in which this occurs are when a set of independent and dependent variable values are both obtained by measurement, then used either (a) to statistically estimate a model (e.g., a regression model) or (b) to calibrate or validate a model that has been constructed based on other means.

Variable Units in ESML

Here are some example units:

- degrees C
- mm mo^-1
- mm
- unitless
- %
- 10^6 gallons d^-1
- m^2/m^2
- no. x 10^3
- mm time^-1
- % of saturation
- gram C m^-2 mo^-1
- 10^6 m^3 time^-1

Variable Temporal Characteristics

Variable Temporal Extent

Identify the earliest and latest dates represented by the source data for this variable. If a variable does not have a temporal extent, enter “not applicable.” All **Predictor:Constant or parameter** variables should be “not applicable.” If no information is available on the dates to which variable data apply, enter "Not reported."

Be aware that different time- or space-varying variables in a given EM may have different temporal extents, and that these collectively determine the extent for any computed response variables and for the EM overall. For example, in one EM (EnviroAtlas - natural filtration), temporal extent for various time- or space-varying predictor variables, determined according to their data sources, was entered as follows: 2010 for land cover and census variables, 2008 for weather variables and 1999 for pollutant pooled mean concentration data (from meta-analysis). Temporal extent for the model per se, and for the computed response variable pollutant load, was therefore entered as the range of all these data, i.e., 1999-2010.

Temporal Grain Size Value

This factor is only applicable for variables with a listed temporal extent that have a value that changes as a function of time and the "grain" of time that is regular (for example, comes from a daily or annual record). Enter the value and unit for regular temporal grain. For example, if variable is based on monthly data, value is "1" and unit is "month."

Variable grain may be coarser than, but can be no finer than, that of the model. For example, if model has a daily time step but variable is from a monthly data set, then variable value probably is defined monthly. However, if a smoothing function is used to create an estimated daily data set, then variable grain in the model is daily. By contrast, if variable observation data set is daily yet model has only a monthly time step, variable grain size cannot be smaller than monthly.

Variable Spatial Characteristics

Variable Spatial Extent Area

Identify the geographic scale of the data for this variable. All **Predictor:Constant or parameter** variables should be “not applicable.” If no information is available on the dates to which variable data apply, enter "Not reported." This category is prefilled with the following options:

- Not applicable
- <1 ha
- 1-10 ha
- 10-100 ha
- 1-10 km2
- 10-100 km^2
- 100-1000 km^2
- 1000-10,000 km^2
- 10,000-100,000 km^2
- 100,000-1,000,000 km^2
- >1,000,000 km^2
- Not reported

Spatial Grain Size Value

Grain is the size of the individual units of observation). For spatial field data, grain size would be the area over which data were averaged to make up one sampling unit. This factor is only applicable for variables with a listed spatial extent area that have a value that changes as a function of location. This category is prefilled with the following options:

- Not applicable
- Length, for linear feature (e.g., stream mile)
- Area, for pixel or radial feature
- Volume, for 3-D feature
- Map scale, for cartographic feature
- Other (specify), for irregular (e.g., stream reach, lake basin)

Spatial Grain Size Value

Enter text description of the shape of the spatial grain size at which the values for the variable are defined. Grain may be irregular; variable observations may be distributed by catchment, by river or estuary reach, by habitat patch, etc. Include all applicable dimensions including a vertical dimension if applicable (e.g., 4 km x 4 km; 12 km x 12 km x 16 km). If grain is a rectangular area, be clear, such as through use of parentheses, whether numeric value represents cell area or cell side length [e.g., "4 km x 4 km grid" or "(4 km)^2 grid," not something ambiguous such as "4 km grid"]. If grain size varies, provide the range (e.g., 10 - 240 ha). If parameterization is distributed by physiographic feature, identify the feature (e.g., "by stream reach," "by reservoir 1 km^2 area and 10-m depth layer"). Specify "NHDplus reach" or "NHDplus catchment" where these are the spatial unit for computations.

Step 3: Draft a Variable Relationship Diagram (Optional)

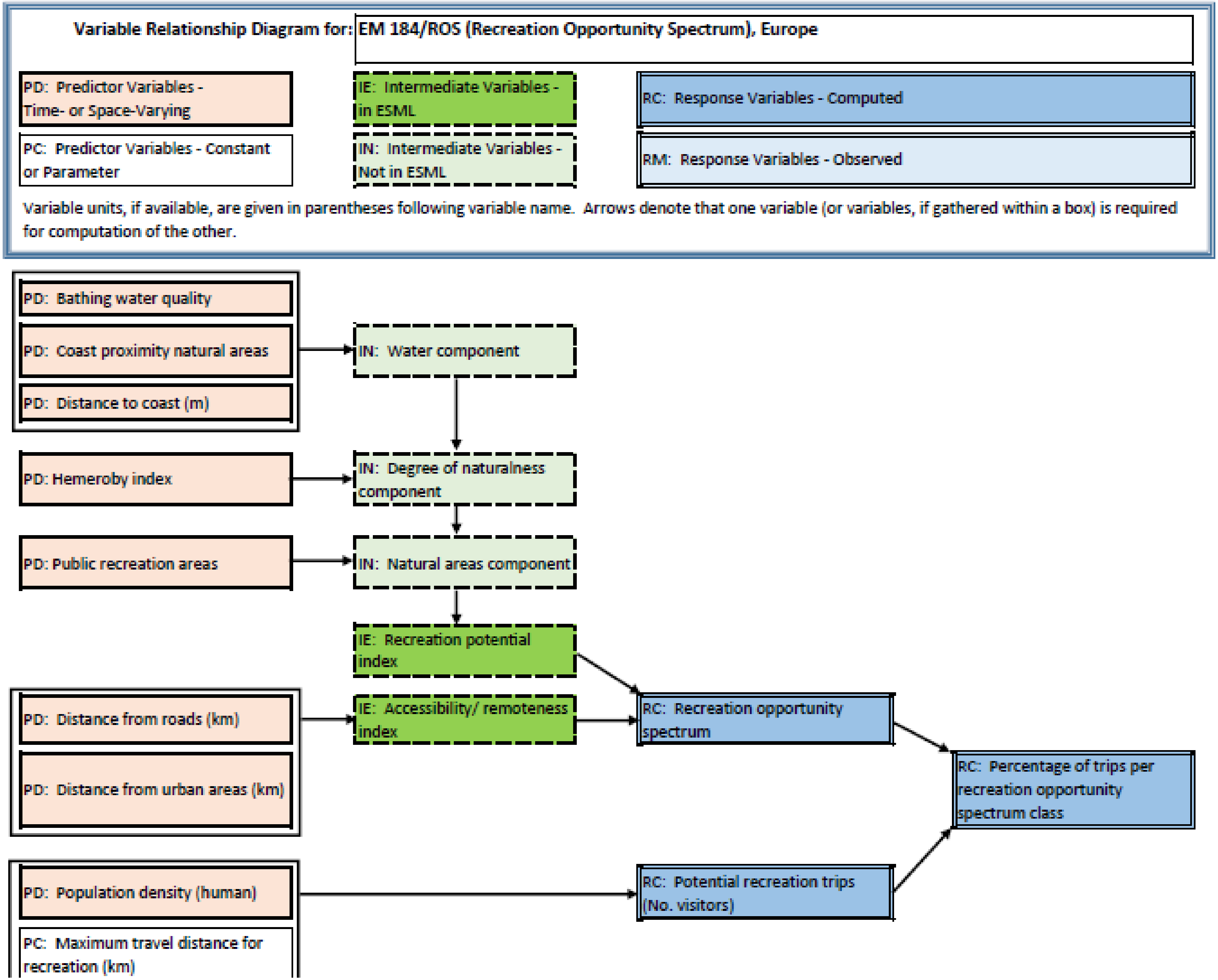
This 3rd step is optional. You are encouraged to sketch a Variable Relationship Diagram to indicate how predictor variables produce response variables. This can be done by hand using pencil or paper or with any software you choose.

In ESML a Variable Relationship Diagram is prepared for each ecological model (EM) showing the types and interrelationships of its variables. Each variable is depicted as a box. Logic flow is left to right (or down). Arrows denote that one variable (or variables, if gathered within a box) is required for computation of the other.

To improve clarity, complex diagrams may show intermediate variables that were computed by the model.

The depicted example diagram, on the next page is for a relatively complex EM, Recreational Opportunity Spectrum (ROS, EM-184). It includes 9 predictor variables (8 Time- or Space-Varying and 1 Constants/Factors). These predictors variously contribute to computation of 5 intermediate and 3 response variables.

A second simpler example, EnviroAtlas - Restorable Wetlands (EM-492) is depicted on the following page. It includes 6 predictor variables (5 Time- or Space-Varying and 1 Constants/Factors). These predictors variously contribute to computation of 2 intermediate and 1 response variables.



Variable Relationship Diagram for: EM 184/ROS (Recreation Opportunity Spectrum), Europe
PD: Predictor Variables - Time- or Space-Varying
PC: Predictor Variables - Constant or Parameter
IE: Intermediate Variables - in ESML
IN: Intermediate Variables - Not in ESML
RC: Response Variables - Computed
RM: Response Variables - Observed
Variable units, if available, are given in parentheses following variable name. Arrows denote that one variable (or variables, if gathered within a box) is required for computation of the other.
PD: Bathing water quality
PD: Coast proximity natural areas
PD: Distance to coast (m)
IN: Water component
PD: Hemeroby index
IN: Degree of naturalness component
PD: Public recreation areas
IN: Natural areas component
IE: Recreation potential index
PD: Distance from roads (km)
PD: Distance from urban areas (km)
IE: Accessibility/ remoteness index
RC: Recreation opportunity spectrum
RC: Percentage of trips per recreation opportunity spectrum class
PD: Population density (human)
PC: Maximum travel distance for recreation (km)
RC: Potential recreation trips (No. visitors)

Variable Relationship Diagram for: EM-492/EnviroAtlas - Restorable wetlands

PD: Predictor Variables - Time- or Space-Varying

IE: Intermediate Variables - in ESML

RC: Response Variables - Computed

PC: Predictor Variables - Constant or Parameter

IN: Intermediate Variables - Not in ESML

RM: Response Variables - Observed

Variable units, if available, are given in brackets following variable name. Arrows denote that one variable (or variables, if gathered within a box) is required for computation of the other.

‡Double dagger denotes a variable whose value is constant with respect to a driving class variable (such as when derived from a lookup table).