Health & Ecological Risk Assessment

Data Quality Scoring System for Microcosm and Mesocosm Studies Used to Derive a Level of Concern for Atrazine

Jeffrey M Giddings,*† David Campana,† Shyam Nair,† and Richard Brain§

†Compliance Services International, Lakewood, Washington, USA
‡Consultant, Idaho Falls, Idaho, USA
§Syngenta Crop Protection, Greensboro, North Carolina, USA

ABSTRACT

The US Environmental Protection Agency (USEPA) has historically used different methods to derive an aquatic level of concern (LoC) for atrazine, though all have generally relied on an expanding set of mesocosm and microcosm (“cosm”) studies for calibration. The database of results from ecological effects studies with atrazine in cosmos now includes 108 data points from 39 studies and forms the basis for assessing atrazine’s potential to impact aquatic plant communities. Inclusion of the appropriate cosm studies and accurate interpretation of each data point—delineated as binary scores of “effect” (effect score 1) or “no effect” (effect score 0) of a specific atrazine exposure profile on plant communities in a single study—is critical to USEPA’s approach to determining the LoC. We reviewed the atrazine cosm studies in detail and carefully interpreted their results in terms of the binary effect scores. The cosm database includes a wide range of experimental systems and study designs, some of which are more relevant to natural plant communities than others. Moreover, the studies vary in the clarity and consistency of their results. We therefore evaluated each study against objective criteria for relevance and reliability to produce a weighting score that can be applied to the effect scores when calculating the LoC. The cosm database includes a wide range of experimental systems and study designs, some of which are more relevant to natural plant communities than others. Moreover, the studies vary in the clarity and consistency of their results. We therefore evaluated each study against objective criteria for relevance and reliability to produce a weighting score that can be applied to the effect scores when calculating the LoC. This approach is useful because studies that are more relevant and reliable have greater influence on the LoC than studies with lower weighting scores. When the current iteration of USEPA’s LoC approach, referred to as the plant assemblage toxicity index (PATI), was calibrated with the weighted cosm data set, the result was a 60-day LoC of 21.2 μg/L. Integr Environ Assess Manag 2018;14:489–497. © 2018 The Authors. Integrated Environmental Assessment and Management published by Wiley Periodicals, Inc. on behalf of Society of Environmental Toxicology & Chemistry (SETAC)

Keywords: Atrazine Mesocosm Data quality Aquatic plants Level of concern

INTRODUCTION

Between 2007 and 2012, the United States Environmental Protection Agency (USEPA) developed a methodology for establishing levels of concern (LoCs) for the herbicide atrazine in freshwater systems, for the protection of plant communities (USEPA 2007b, 2009a, 2012a, 2016; Erickson 2012). The methodology used 2 exposure–response models, the plant assemblage toxicity index (PATI) and the comprehensive aquatic systems model (CASM), each calibrated against a mesocosm and microcosm (“cosm”) data set comprising 87 unique exposure profiles (chemographs) and corresponding binary effect scores (1 for effect, 0 for no effect). The 2 models use radically different approaches: PATI is essentially an average plant species exposure–response curve, while CASM is a mechanistic bioenergetics model with multiple simulated species at each of several trophic levels. CASM was specifically parameterized to represent the toxicity of atrazine, and the parameterized model was designated CASM_{ATZ}. Both models translate an individual chemograph into an estimated cumulative effect endpoint, and a logistic regression is then applied to plots of effect endpoint versus binary cosm effect score to arrive at an LoC calculated as the 50th percentile of the curve. The 60-day time-weighted average concentration that would result in the LoC, termed the concentration equivalent LoC (CE-LoC), is then calculated. Beginning in 2009, USEPA focused on PATI rather than CASM_{ATZ} for LoC determination (USEPA 2009a, 2012a, 2016). However, model-based LoC derivation is highly sensitive to, and influenced by, the interpreted scoring of the calibrating cosm data set. Based on the PATI model, USEPA most recently calculated a preliminary CE-LoC of 3.4 μg/L (USEPA 2016).

The large number of published cosm studies with atrazine represent a unique database for the establishment of an LoC (USEPA 2003, 2007b, 2009a, 2012a, 2016; Erickson 2012; Giddings et al. 2005; Solomon et al. 1996). These community-level studies potentially represent the most environmentally
realistic investigations of the responses to atrazine exposures and therefore provide the best available ecologically relevant endpoints, consistent with the goal of protecting aquatic primary producer community structure and function. After USEPA’s publication of the initial atrazine cosm database (USEPA 2003), detailed investigation revealed that some of the critical studies were unreliable and that others had been misinterpreted (Giddings et al. 2005). These conclusions were corroborated by several meetings of the Federal Insecticide, Fungicide and Rodenticide Act (FIFRA) Scientific Advisory Panel (SAP) (USEPA 2007a, 2009b, 2012b), but some of the SAP’s most significant recommendations were not reflected in USEPA’s most recent iteration of the database (USEPA 2016).

Because the cosm studies varied greatly in their experimental designs and the consistency of their results, we developed a set of criteria for evaluating the relevance and reliability of each data point, and incorporated the data quality weights into the derivation of the LoC. Recent publications by Suter et al. (2017a, 2017b) examined weight-of-evidence (WoE) approaches to inferring qualities and quantities in risk assessment, and Hall et al. (2017) emphasized the importance of weighting individual data points within a line of evidence. Moermond et al. (2017) and Rudén et al. (2017) presented detailed criteria for data relevance and reliability, respectively, and Van Der Kraak et al. (2014) applied such a data weighting system to evaluation of atrazine toxicity to amphibians.

Moore et al. (2017) applied a WoE approach to the derivation of the atrazine LoC. Four lines of evidence were considered: cosm data with PATI, cosm data with CASM ATZ, visual examination of cosm data alone, and standard methods for deriving water quality criteria (Stephan et al. 1985). Each line of evidence was weighted on the basis of characteristics of each method, and the weights of each line of evidence were combined to derive an overall (weighted) LoC. Moore et al. used a data weighting scheme similar to ours to evaluate the studies identified as questionable by the 2007, 2009, and 2012 SAPs (USEPA 2007a, 2009b, 2012b), scoring each study for relevance (yes/no) and reliability (acceptable, supplemental, unacceptable). The current paper presents an extension of the Moore et al. (2017) analysis. We developed data evaluation criteria for individual cosm studies similar to (but independently from) Moore et al. (2017). We then applied the weighting scores for individual data points to CE-LoC_{PATI}, CE-LoC_{CASM-ATZ}, and visual LoC derivations to observe the effect of the data evaluation on the derived LoC.

**METHODS**

**Atrazine cosm database**

The atrazine cosm database was first compiled by USEPA as part of the Interim Reregistration Eligibility Decision for atrazine (USEPA 2003) and evolved in response to new data as well as recommendations from the FIFRA Scientific Advisory Panel (SAP) (USEPA 2007a, 2009b, 2012b). In constructing the database, a “study” was defined as a unique experimental system (test system design, source of components, exposure regime, experimental conditions) with controls plus one or more treatment groups. A data point was defined by the response of 1 plant community (phytoplankton, periphyton, or macrophytes) in 1 treatment group (exposure regime) in 1 study. The exposure regime associated with each data point was originally characterized by initial atrazine concentration and duration (days between the initial atrazine application and the end of the observation period). In 2009, the cosm exposure regimes were recharacterized by constructing the daily time series of measured or inferred atrazine concentrations in the water column (USEPA 2009a). The observed response of each plant community was initially characterized with a 5-point scale ranging from “no effect” to “pronounced effect without return to control levels for more than 56 d” (Brock et al. 2000). The 5-point effect scores were later replaced by a binary scoring system in which points corresponding to Brock scores of 3 through 5 were assigned a score of 1 (“effect”) and points corresponding to Brock scores of 1 and 2 were assigned a score of 0 (“no effect”).

Between 2003 and 2016, data points were added from new studies, while some data points were removed in response to suggestions from the SAP and others. The most recent version (USEPA 2016) has inconsistencies regarding separation of data points for multiple studies and for different plant communities—in some cases studies or communities are represented separately, while in other cases they are combined. In our analysis, we split data points as appropriate. Accounting for the splits, the database currently contains 108 data points (Supplemental Table SI-1) from 39 studies (Supplemental Table SI-2).

**Evaluation of effects**

We evaluated the responses of the cosmats to atrazine on the basis of published reviews and guidelines (Brock et al. 2000; Campbell et al. 1999; De Jong et al. 2008; Giddings et al. 2002). Effects on community-level attributes (e.g., diversity, productivity, biomass) were given priority, but effects on individual populations were also considered. As discussed above, the responses of cosm phytoplankton, periphyton, or macrophytes to each exposure regime in each reported study were scored as 1 (effect) or 0 (no effect) based on consideration of the following:

- Were effects “slight” or “pronounced”?
- Did statistically significant differences from controls occur on at least 2 consecutive sampling dates?
- Was the onset of effects consistent with the rapid mode of action of atrazine?
- Were clear exposure–response relationships observed?
- Were effects transient, and if so, when did recovery occur?

The scoring results were consistent with USEPA’s (2016) for most data points. However, 14 data points were rescored from “1” to “0” for reasons presented previously (Giddings 2012; Giddings et al. 2005; Solomon et al. 1996) and affirmed
by the SAP (USEPA 2009b, 2012b). Effect scores for each data point are provided in Supplemental Table SI-1.

Evaluation of data quality

Due to the unique and important role that cosm studies have played in the history of atrazine risk assessment (Brock et al. 2000; USEPA 2016; Giddings et al. 2005; Solomon et al. 1996), there has long been scientific discussion and debate about which data are most reliable and which should be accepted as the basis for regulatory decisions. By applying a set of objective evaluation criteria to the atrazine cosm studies, higher importance can be assigned to studies of better quality in a WoE analysis, while minimizing the contribution of poor or irrelevant studies to inferences derived from the assessment.

We developed a set of 10 criteria believed to be useful in judging the overall quality of a cosm study (Table 1). While these criteria were developed with the atrazine cosms in mind, for the most part they are not specific to atrazine or cosm analysis; they are consistent with several recent publications about approaches to WoE analysis (Hall et al. 2017; Moermond et al. 2017; Moore et al. 2017; Rudén et al. 2017; Suter et al. 2017a, 2017b). Four criteria were designed to evaluate study relevance, and 6 criteria addressed study reliability.

Relevance. A relevant atrazine cosm study is one designed to mimic exposure from drift and runoff or erosion in a natural environment such as a small pond or stream. Factors considered are the physical structure of the test system, the size of the test system, ecological complexity, and the realism of the exposure regime.

The real ecosphere is complex and varies continuously in space and time. Mesocosm studies enable researchers to capture some of this complexity, but studies can and do vary in the degree to which the ecosystems they represent are aligned with the spatial and temporal scales and ecological complexity of the risk assessment conceptual model. Our relevance criteria are based on the conceptual spatial scale of a small (first to third order) stream or a small (e.g., 1 ha) pond. We presume that small lentic and lotic systems such as these are the most vulnerable to pesticide exposure. Based on the

| Criterion                                      | Score = 0                              | Score = 1                                      | Score = 2                                      |
|------------------------------------------------|----------------------------------------|-----------------------------------------------|-----------------------------------------------|
| **Relevance**                                  |                                        |                                               |                                               |
| Physical structure of test system              | Highly artificial                      | Incorporates some natural aspects             | Closely simulates natural pond or stream       |
| Size of test system                            | ≤1 m³                                  | 1–10 m³                                      | >10 m³                                        |
| Ecological complexity                          | Highly simplified                      | Multiple trophic levels                       | Complex communities                            |
| Realism of exposure regime                     | Constant exposure                      | Unrealistic time-varying exposure             | Realistic pulsed exposure                      |
| **Reliability**                                |                                        |                                               |                                               |
| Number of replicates                           | No replication                         | Two replicates                                | Three or more replicates                      |
| Atrazine analysis                              | No analysis of exposure concentrations | Analysis of initial concentrations            | Repeated analysis over time                    |
| Consistency of response among replicates       | High variability among replicates      | Low to moderate variability among replicates for most endpoints | Low variability among replicates               |
| Consistency of exposure response               | Inconsistent exposure response         | Generally consistent exposure response        | Consistent monotonic exposure response         |
| Consistency of response over time              | Delayed response or response varies over time unrelated to treatment regime | Response not clearly linked to time of atrazine exposure | Immediate response to treatment; smooth temporal trends |
| Influence of confounding factors               | Response confounded by intentional or inadvertent influence of other factors (e.g., other chemicals, nutrient supply, grazing, temperature, stress) | Confounding factors may have influenced response but response (or lack of response) to atrazine still evident | Response clearly attributable to atrazine exposure, unrelated to other factors |
mode of action of atrazine, the temporal scale of interest is 
assumed to be weeks to months. “Level of complexity” is 
evaluated relative to both the physical and ecological complexity of a healthy natural ecosystem. Greater complexity 
results in higher scores for relevance, but greater complexity may also add variability and make interpretation 
more difficult, reducing scores for reliability.

Physical structure of the test system. Especially when the 
stressor of concern is an herbicide, it is important to provide 
physical habitat to support natural or naturally derived 
phytoplankton, periphyton, and/or macrophyte communities. 
The more closely the physical structure of the test system 
resembles a small natural pond or stream, the better it will match 
the assessment endpoints. If the test system is highly simplified 
or artificial, observed ecological responses to chemical expo-
sure will need greater extrapolation to natural situations.

Size of the test system. Gauged against the standard of a 
small natural pond or stream, the realism—and therefore the 
relevance—of the cosm environment is largely a function of 
the size of the test system. Cosms vary greatly in size, but 
generally fall into 3 categories: bench scale (≤ 1 m²), small 
outdoor (≤ 10 m²), and larger outdoor ponds or streams. Smaller systems may be adequate for phytoplankton studies; 
larger systems would be needed for macrophyte studies.

Ecological complexity. Ecological complexity is a desirable 
aspect of cosm design for several reasons:

- A wide range of taxa is represented;
- Ecological complexity enables direct measurement of 
  community function and structure;
- Ecological complexity enables observation of population 
  and community responses in a realistic ecological context 
  (e.g., nutrient supply, grazers, competitors).

While the scoring criteria for ecological complexity (Table 1) 
may appear to be somewhat subjective, most actual cosm 
studies tend to fall clearly into 1 of the 3 categories. “Highly 
simplified” cosms may be assemblages created by combining 
single species, or systems that include only a single trophic 
level. At the other end of the complexity spectrum are natural 
or naturally derived communities with all major ecosystem 
components represented. Between these extremes are 
systems whose simplifications may compromise ecological 
complexity in favor of other experimental objectives.

Realism of exposure regime. Realistic atrazine exposure 
regimes are exemplified by the chemographs measured 
during the atrazine midwestern monitoring program (USEPA 
2007b) and vary from single to repeated concentration pulses 
followed by disappearance from the water column. The most 
realistic (relevant) cosm exposure regimes, especially for lotic 
systems, are single pulses or series of pulses simulating 
episodic entry of spray drift or field runoff. If concentrations 
are artificially manipulated (e.g., step-wise increases over 
time [Detenbeck et al. 1996]), or the test system lacks one or 
more critical environmental fate processes (e.g., removal of 
chemical from the water by periphyton and macrophytes), the 
exposure regime is likely to be less realistic and therefore the 
study will be intermediate in relevance. Cosm studies with 
continuous atrazine concentrations are not representative of 
real-world exposure and receive the lowest rating for realism 
of exposure regime.

Reliability. These criteria evaluate the studies in terms of the 
reliability of inferences (effect/no effect decisions on individ-
ual data points) that can be drawn from them. Two criteria are 
objective (number of replicates and atrazine analysis), while 
the other 4 are more subjective, examining the consistency of 
responses across replicates, concentrations, and time as well 
as the influence of confounding factors.

Number of replicates. Especially in this case where the 
inference is based on a comparison of treated cosms with 
controls, the power of the test is a function of the number of 
treated and control replicates. A study without replicates is 
extremely unreliable; duplicates are the minimum for even the 
weakest subjective inferences; triplicates (or more) are standard.

Atrazine analysis. Data evaluators are often ruthless about 
excluding exposure–response data that are unaccompanied 
by analytical confirmation of the exposure concentrations. 
However, in the case of cosm studies, reliable information can 
be obtained even without chemical analysis. Therefore, our 
scoring system did not reject studies outright if there was no 
atrazine analysis, but they were given the lowest score on this 
criterion. Analysis of initial concentrations in the exposure 
water was the minimum requirement for an intermediate 

criterion, Analysis of oginal concentrations, while 
the highest score was reserved for studies with a 
series of atrazine analyses over a relevant period.

Consistency of response among replicates. The increased 
test power associated with a high number of replicates is 
useful because the consistency (or lack thereof) among 
replicates within a treatment group can inform the degree of 
confidence or certainty associated with a result. For example, 
if periphyton mass decreases over time in one treated 
replicate but remains constant in another, it is not clear 
whether an atrazine effect has occurred.

Consistency of exposure response. In the case of atrazine, it 
is reasonable to assume that the severity of a true effect will 
be a monotonic function of the dose administered. The 
consistency of the exposure response informs whether an 
observed change (e.g., periphyton decreasing over time) is 
due to atrazine exposure. If the trend across treatment levels 
within a study is not monotonic, the observed response may 
be due to causes other than atrazine. This criterion is helpful in 
distinguishing causation from correlation.

Consistency of response over time. Cosms are inherently 
more complex systems than single-species experiments, and
it is common for certain species or groups to change over time due to factors unrelated to the treatment. The consistency of response over time, relative to atrazine’s mode of action, can be used to judge whether an observed trend is likely to be an atrazine effect. Atrazine would be expected to have immediate but reversible effects on primary productivity, with the possibility of later effects on population density or biomass due to cumulative effects on productivity. Changes in primary productivity that are observed only days or weeks after atrazine treatment are unlikely to be caused by atrazine.

Influence of confounding factors. Some experiments are subject to circumstances that may compromise some or all the results observed in the studies. Under the umbrella of “confounding factors,” these circumstances might include, for example, unplanned perturbations to some or all replicates, simultaneous exposure to other chemicals, or the uncontrolled introduction of a selective grazer.

Data evaluation process

Each study was rated with a data quality score of 0, 1, or 2 for each of the criteria summarized in Table 1. Scores were assigned at the study level and then applied to each data point. In 2 cases, different data points from an individual study were assigned different weights; see details in Supplemental Table S1-3). Relevance was characterized by the sum of the scores for the 4 relevance criteria (maximum score = 8). Reliability was characterized by the sum of scores for the 6 reliability criteria (maximum score = 12). The overall score was calculated as the sum of relevance and reliability scores (maximum score = 20). Each data quality score was expressed as a decimal fraction of the corresponding maximum score.

LoC_{PATI} and LoC_{CASM-ATZ} derivation with weighted data points

The PATI model and methods used in this analysis to estimate LoCs are presented in detail in Nair and Brain (2012) and Erickson (2012). Analysis of weighted data points required multiple assessments of each data point according to its weight. It was not possible to assess larger data sets without making changes within the source code. USEPA’s 2016 version of PATI could not be used because the source code was not available at the time of this analysis. Consequently, the 2012 version of the PATI model defined in Erickson (2012) and Nair and Brain (2012) was used instead.

LoCs were also estimated with the CASM_{ATZ} model (Bartell et al. 2013; Nair et al. 2015). CASM_{ATZ} estimates the potential toxic effects of atrazine on populations of aquatic plants and consumers in a generic lower-order midwestern stream as the deviation in the maximum 60-day average SSI from a reference, pristine, stream. CASM_{ATZ} simulates the daily production of 20 periphyton, 6 aquatic vascular plant species, and 17 functionally defined species of zooplankton, benthic invertebrates, bacteria, and fish representing the consumer community. SSIs are estimated from daily values of population biomass, calculated as nonlinear functions of population bioenergetics, physical and chemical environmental parameters, grazing or predator-prey interactions, and population-specific direct and indirect responses to atrazine. Estimates of deviation in the maximum 60-day average SSI and the binary cosmos score for each Cosmos are then used to estimate the LoC_{CASM-ATZ} as in Erickson (2012).

As PATI and CASM_{ATZ} generate different measures of effect, the LoCs need to be transformed to a uniform measurable quantity for meaningful application in the real world. As each approach used maximum estimated 60-day LoCs, the 60-day constant exposure concentration (during a fixed growing period starting at Julian day 105) that would result in the same value of LoC as predicted by PATI or CASM_{ATZ} was calculated. This value was called the 60-day CE-LoC (µg/L).

RESULTS AND DISCUSSION

How our evaluation system relates to others

Rudén et al. (2017) presented an approach to evaluating data relevance, acknowledging there is no universally accepted definition of study relevance. They considered exposure relevance issues to include the test substance, its concentration and quantity of exposure, and exposure
dynamics. Generally, the studies in the database used technical grade atrazine active ingredient, thus precluding concern about the relevance of the test substance. However, there was considerable variation in the other factors listed by Rudén et al. (2017), and these are captured in the “realism of the exposure regime” criterion in our system. The biological relevance topics addressed by Rudén et al. (2017) included selection of test species as surrogates for species of concern in the environment, the physiology of the test organism, and the vulnerability of test organisms in higher-tier community-level assessments. These criteria (particularly the community-level observations that address species recovery) parallel our “ecological complexity” criterion.

Moermond et al. (2017) reviewed 9 toxicity data scoring systems against a set of criteria developed to compare methods of assessing data reliability. Their leading conclusion was that a “systematic and transparent assessment method” is critical to assessing reliability. Our approach (Table 1) was developed to minimize subjectivity. The criteria corresponding to number of replicates and analysis of atrazine concentration over time were entirely objective. Determinations of consistency among replicates, across treatment levels, or over time were more subjective, but entirely transparent. The additional criterion of the “influence of confounding factors” was intended to provide weight to studies with the clearest causal relationships. Moermond et al. (2017) also concluded that studies not conducted in compliance with strict laboratory guidelines should not be excluded on this basis. Similarly, our approach did not consider test guidelines or Good Laboratory Practice standards in scoring these studies because regulatory guidance for cosm studies is lacking. The merits of each study were judged according to the procedures and execution of each experiment for each endpoint.

Moore et al. (2017) applied their scoring rubric only to the atrazine cosm studies that were identified as questionable by the SAP (USEPA 2009b, 2012b), plus 5 new studies. Data relevance was evaluated with 5 binary criteria, while data quality was scored quantitatively with 10 criteria scored 1–3. Unlike our approach, Moore et al. (2017) summarized the results of relevance and quality independently, grading the former as “relevant” or “not relevant” and the latter (based on numerical scoring) as “acceptable,” “supplemental,” or “unacceptable.” The criteria for relevance included:

- Was the study relevant to communities of aquatic plants? This criterion is encompassed by our “ecological complexity” criterion.
- Was atrazine the only active ingredient to which test organisms were exposed? This was a fundamental consideration in identification of data points, so all data used in our analysis met this criterion.
- Were the test endpoints direct measures of community-level effects for aquatic plants? This criterion corresponded, in part, to the “ecological complexity” criterion in our scheme. Moore et al. (2017) rated studies “not relevant” if the measured responses addressed only selected species rather than community-level attributes; our scheme accepted data for individual species but scored them lower than data representing full communities.
- Was the exposure route in the study relevant to what is expected in the environment? This criterion corresponds directly to our “realism of exposure” criterion.
- Was a recovery phase included? Moore et al. (2017) rated studies “not relevant” if a recovery phase (postexposure observation period) was not included. We did not consider a recovery phase to be an essential component of cosm study design, so many of the studies rejected by Moore et al. (2017) were retained in our analysis.

Most of the reliability criteria used by Moore et al. (2017) corresponded to those in our evaluation system, with some exceptions. We did not consider whether studies were conducted according to recognized international standards.

| Data set         | 60-day CE-LoC\textsubscript{Pati} (\(\mu g/L\)) | 60-day CE-LoC\textsubscript{CASM\textsubscript{ATZ}} (\(\mu g/L\)) |
|------------------|-----------------------------------------------|---------------------------------------------------------------|
|                  | Unweighted | Weighted | Unweighted | Weighted |
| EPA (2016)       | 2.71       | 4.18     | 7.72       | 8.33     |
| This analysis    | 18.6       | 21.2     | 16.8       | 18.6     |

\*CE-LoC = concentration equivalent level of concern; PATI = plant assemblage toxicity index; CASM = comprehensive aquatic systems model; ATZ = atrazine.

Figure 3. Effect scores as a function of initial atrazine concentration and exposure duration, using data with overall evaluation scores of 0.7 or greater. Open squares: No effect. Filled circles: Effect. Open squares overlap with filled circles for 3 data points.
because few such standards exist for cosms and those that do (e.g., CLASSIC, Giddings et al. 2002) emphasize the need to adapt study designs to specific risk assessment questions. Similarly, as mentioned above, compliance with Good Laboratory Practice standards was not a criterion in our system. The other reliability criteria presented by Moore et al. (2017) were considered explicitly or implicitly in our criteria.

**Evaluation scores**

Overall scores for each data point are summarized in Supplemental Table SI-1, and scores for individual criteria for each study are summarized in Supplemental Table SI-2. The detailed evaluations of individual studies and data points are presented in Supplemental Table SI-3.

The distribution of overall evaluation scores for individual data points is shown in Figure 1. The highest-rated study (King et al. 2016) received an overall score of 0.95. At the other end of the range, 2 studies were scored 0.25. Only 2 studies (Baxter et al. 2011; King et al. 2016) were rated 0.85 or higher overall. King et al. (2016) was the only study to rate a 1.0 for study relevance, indicating an experimental system best simulating a natural environment that could be subjected to atrazine inputs. Predictably, larger test systems also tended to have high ecological complexity; of 11 studies that were scored 2 for test system size, only 1 (McGregor et al. 2008) was scored less than 2 for ecological complexity.

Eight studies (represented by 35 data points) received an overall rating of 0.7 or higher. These were studies rated highest for relevance and/or reliability (Figure 2). The best-rated studies were distinguished by higher ratings for the 3 criteria related to consistency, which tended to be the lowest-rated reliability criteria across all studies. The 5 highest-rated studies averaged scores of 0.73 for consistency of response among replicates, consistency of exposure response, and consistency over time. The next 9 highest studies (those with overall scores ≥0.55) averaged consistency scores of 0.48, and consistency scores continued to trend downward for the remaining studies.

**Impact of data quality scores on LoC**

The 60-day CE-LoC\textsubscript{PATI} and CE-LoC\textsubscript{CASM-ATZ} were estimated from the 2 cosm sets (USEPA 2016 and our revised data set, both presented in Supplemental Table SI-1) with data weighted or unweighted based on the data quality scores. Results are summarized in Table 2. The CE-LoC\textsubscript{PATI} based on the unweighted USEPA 2016 cosm set was 2.71 μg/L. (Note that this value is slightly different than the value of 3.4 μg/L generated by USEPA (2016), likely due to differences between versions of the PATI model as discussed in the Methods section.) With the same data, CE-LoC\textsubscript{CASM-ATZ} was 7.72 μg/L, nearly 3 times greater than CE-LoC\textsubscript{PATI}, showing the significant influence of the mechanistic model predictions on the logistic exposure–response curve. In the LoC calculations under both models, a logistic curve is fitted to the binary score (0 or 1) for each data point (USEPA 2012a). This statistical process is highly susceptible to extreme values (e.g., studies associating a binary effect score of 1 with extremely low exposure concentrations), which can result in a skewed regression and accordingly a biased LoC estimation. Moreover, the CE-LoC\textsubscript{PATI} increased by 54% and CE-LoC\textsubscript{CASM-ATZ} increased by 8% when data quality scores were factored into the analysis based on the USEPA 2016 data set.

With the revised data set, the unweighted CE-LoC\textsubscript{PATI} and CE-LoC\textsubscript{CASM-ATZ} were similar (18.6 and 16.8 μg/L, respectively) and more than twice as high as the CE-LoC\textsubscript{CASM-ATZ} for the USEPA data set. When data quality scores were factored into the analysis, the CE-LoC\textsubscript{PATI} and CE-LoC\textsubscript{CASM-ATZ} increased by 14% and 11% to 21.2 and 18.6 μg/L, respectively.

Both PATI and CASM\textsubscript{ATZ} generated 60-day CE-LoC values of about 20 μg/L with the revised data set. For PATI, the nonmechanistic model, selection of the USEPA 2016 data set or the revised data set led to a 5- to 7-fold difference in CE-LoC. With the mechanistic CASM\textsubscript{ATZ} model, results for the 2 datasets differed only about 2-fold. Data quality scoring had a small but measurable effect on CE-LoC values derived with both models and both data sets.

A more direct approach to deriving an LoC from the atrazine cosm data is a “simple and transparent time and concentration-based LoC,” as recommended by the 2009 SAP (USEPA 2009b). Throughout many versions of the atrazine cosm database, effect or no effect results were arrayed on axes representing initial atrazine concentration and duration of exposure; most recently, Moore et al. (2017) used this graphical presentation as 1 of the 4 lines of evidence they considered. To incorporate the results of the data quality evaluation into this visual analysis, we focused on the highest quality studies. Figure 3 shows the distribution of effect or no effect results for the 35 data points with overall scores of 0.7 or greater. The lowest atrazine concentration to produce an effect in these studies was 50 μg/L (initial concentration), regardless of exposure duration.

**CONCLUSIONS**

- The cosm evaluation system succeeded in objectively distinguishing studies of high quality such as Baxter et al. (2011) and King et al. (2016) from those of lesser quality. With appropriate minor modification, the evaluation system could be adapted for other chemical classes.
- When the data quality weighting results were incorporated into the PATI and CASM\textsubscript{ATZ} LoC models, the atrazine LoC increased by 8% to 12% (or more, for preliminary CE-LoC\textsubscript{PATI} with the USEPA 2016 data set). This indicates that unweighted LoC values are disproportionately affected by data from studies of lower quality.
- The 60-day CE-LoC values based on the revised data set, weighted for relevance and reliability, were 21.2 μg/L with the PATI model and 18.6 μg/L with the CASM\textsubscript{ATZ} model. These values were similar to those derived by Moore et al. (2017).
The best studies (i.e., those with overall data quality scores of 0.7 or greater) indicated that atrazine did not affect plant communities when initial atrazine concentrations were less than 50 µg/L, regardless of exposure duration.

Data evaluation is an integral part of WoE analysis.

Acknowledgment—This work was supported by Syngenta Crop Protection, Greensboro, NC, USA. R Brain is an employee of this company and primarily conceived the conceptual approach. Study reviews were independently conducted by J Giddings and D Campana, and LoC derivation was independently conducted by S Nair.

Disclaimer—The peer review process for this article was managed without the involvement of R Brain.

Data Accessibility—All supporting data and tools cited in this paper are available from the corresponding author, Jeffrey Giddings, at jgiddings@complianceservices.com.

SUPPLEMENTAL DATA
Table SI-1. Database summary (by data point, including ID#, effect score, relevance, reliability, overall weight).

Table SI-2. Database summary (by study, individual criterion scores).

Table SI-3. Evaluation records for each study/data point.

REFERENCES
Bartell S, Brain R, Hendley P, Nair S. 2013. Modeling the potential effects of atrazine on aquatic plant communities in midwestern streams. *Environ Toxicol Chem* 32:2402–2411.

Baxter L, Moore D, Sibley P, Solomon K, Hanson M. 2011. Atrazine does not affect algal biomass or snail populations in microcosm communities at environmentally relevant concentrations. *Environ Toxicol Chem* 30:1689–1696.

Brock T, Lahr J, van den Brink P. 2000. Ecological risks of pesticides in freshwater ecosystems. Part 1. Herbicides. Wageningen (NL): Alterra, Report 088.

Campbell P, Arnold D, Brock T, Grandy N, Heger W, Heimbach F, Maund S, Strelke M. 1999. Guidance document on higher-tier aquatic risk assessment for pesticides (HARAP), from the SETAC-Europe/OECD/EC workshop, Lacanau Ocean, France. Brussels (BE): Society of Environmental Toxicology and Chemistry Europe (SETAC-Europe).

De Jong F, Brock T, Foekema E, Leeuwvang P. 2008. Guidance for summarizing and evaluating aquatic micro- and mesocosm studies. Bilthoven (NL): National Institute for Public Health and the Environment, RIVM Report 601 506009/2008.

Detenbeck N, Hermanutz R, Allen K, Swift M. 1996. Fate and effects of the herbicide atrazine in flow-through wetland mesocosms. *Environ Toxicol Chem* 15:937–946.

Erickson R. 2012. Proposed methodology for specifying atrazine levels of concern for protection of plant communities in freshwater ecosystems. Duluth (MN): US Environmental Protection Agency, Report EPA/600/R-12/019.

Giddings J. 2012. Atrazine review of mesocosm and microcosm studies evaluation. Lakewood (WA): Compliance Services International, Report No. 12709.

Giddings J, Anderson T, Hall L, Hosmer A, Kendall R, Richards R, Solomon K, Williams W. 2005. Atrazine in North American surface waters: A probabilistic aquatic ecological risk assessment. Pensacola (FL): Society of Environmental Toxicology and Chemistry (SETAC). 432 p.

Giddings J, Heger W, Brock T, Heimbach F, Maund S, Norman S, Ratte H, Schäfers C, Strelke M. 2002. Community-level aquatic system studies – interpretation criteria (CLASSIC), Fraunhofer Institute, Schalldenberg, Germany. Pensacola (FL): Society of Environmental Toxicology and Chemistry (SETAC). 60 p.

HALL A, Belanger S, Guiney P, Galay-Burgos M, Maack G, Stubblefield W, Martin O. 2017. New approach to weight-of-evidence assessment of ecotoxicological effects in regulatory decision-making. *Integr Environ Assess Manag* 13:573–579.

King R, Brain R, Back J, Becker C, Wright M, Djomte V, Scott W, Virgil S, Brooks B, Hosmer A, Chamblias C. 2017. Effects of pulsed atrazine exposures on autotrophic community structure, biomass, and production in field-based stream mesocosms. *Environ Toxicol Chem* 35:660–675.

McGregor E, Solomon K, Hanson M. 2008. Effects of planting system design on the toxicological sensitivity of Myriophyllum spicatum and Elodea canadensis to atrazine. *Chemosphere* 73:249–260.

Moermond C, Beasley A, Breton R, Junghans M, Laskowski R, Solomon K, Zahner H. 2017. Assessing the reliability of ecotoxicological studies: An overview of current needs and approaches. *Integr Environ Assess Manag* 13:640–651.

Moore D, Greer C, Manning G, Wooding K, Beckett K, Brain R, Marshall G. 2017. A weight-of-evidence approach for deriving a level of concern for atrazine that is protective of aquatic plant communities. *Integr Environ Assess Manag* 13:688–701.

Nair S, Bartell S, Brain R. 2015. A comparative study of the modeled effects of atrazine on aquatic plant communities in midwestern streams. *Environ Toxicol Chem* 34:2590–2602.

Nair S, Brain R. 2012. Description and sensitivity analysis of EPA’s proposed methodology for specifying levels of concern for protection of plant communities in freshwater ecosystems using the PAT1 model. Greensboro (NC): Syngenta Crop Protection.

Rudén C, Adams J, Agerstrand M, Brock T, Poulsen V, Schlekat C, Wheeler J, Henry T. 2017. Assessing the relevance of ecotoxicological studies for regulatory decision making. *Integr Environ Assess Manag* 13:652–663.

Solomon K, Baker D, Richards R, Dixon K, Klaine S, La Point T, Kendall R, Weisskopf C, Giddings J, Giesy J, Hall LW Jr, Williams WM. 1996. Ecological risk assessment of atrazine in North American surface waters. *Environ Toxicol Chem* 15:31–76.

Stephan C, Mount D, Hansen D, Gentile J, Chapman G, Brungs W. 1998. Guidelines for deriving numerical national water quality criteria for the protection of aquatic organisms and their uses. Washington (DC): US Environmental Protection Agency.

Suter G, Cormier S, Barron M. 2017a. A weight of evidence framework for environmental assessments: Inferring quantities. *Integr Environ Assess Manag* 13:1038–1044.

Suter G, Cormier S, Barron M. 2017b. A weight of evidence framework for environmental assessments: Inferring quantities. *Integr Environ Assess Manag* 13:1045–1051.

[USEPA] US Environmental Protection Agency. 2003. Interim reregistration eligibility decision (RED): Atrazine. Washington (DC): US Environmental Protection Agency, Report EPA 738-R-04-004.

[USEPA] US Environmental Protection Agency. 2007a. FIFRA Scientific Advisory Panel, Notice of public meeting on the ecological significance of atrazine stream-water concentrations using a statistically-designed monitoring program. Washington (DC): US Environmental Protection Agency.

[USEPA] US Environmental Protection Agency. 2007b. Preliminary interpretation of the ecological significance of atrazine stream-water concentrations using a statistically-designed monitoring program. Washington (DC): US Environmental Protection Agency.

[USEPA] US Environmental Protection Agency. 2009a. The ecological significance of atrazine effects on primary producers in surface water streams in the corn and sorghum growing region of the United States (Part I). Washington (DC): US Environmental Protection Agency.

[USEPA] US Environmental Protection Agency. 2009b. SAP Minutes No. 2009-14, 2018—JM Giddings et al.

Zahner H. 2017. Assessing the reliability of ecotoxicological studies: An overview of current needs and approaches. *Integr Environ Assess Manag* 13:688–701.
[USEPA] US Environmental Protection Agency. 2012a. Meeting of the FIFRA Scientific Advisory Panel on the problem formulation for the environmental fate and ecological risk assessment for atrazine. Washington (DC): US Environmental Protection Agency.

[USEPA] US Environmental Protection Agency. 2012b. SAP Minutes No. 2012-05. A set of scientific issues being considered by the Environmental Protection Agency regarding: Problem formulation for the reassessment of ecological risks from the use of atrazine. Washington (DC): US Environmental Protection Agency.

[USEPA] US Environmental Protection Agency. 2016. Refined ecological risk assessment for atrazine. Washington (DC): US Environmental Protection Agency.

Van Der Kraak G, Hosmer A, Hanson M, Kloas W, Solomon K. 2014. Effects of atrazine in fish, amphibians, and reptiles: An analysis based on quantitative weight of evidence. Crit Rev Toxicol 44:1–66.