TITLE: Evaluating the environmental parameters that determine aerobic biodegradation half-lives of pesticides in soil with a multivariable approach

AUTHORS: Yuxin Wang[1], Adelene Lai[2,3], Diogo Latino[2], Kathrin Fenner[2,3,4], and Damian E. Helbling[1]

AFFILIATION: [1]School of Civil and Environmental Engineering, Cornell University, Ithaca, NY, USA; [2]EAWAG, Überlandstrasse 133, 8600 Dübendorf, Switzerland; [3]Institute of Biogeochemistry and Pollutant Dynamics, ETH Zürich, 8092 Zürich, Switzerland; [4]Department of Chemistry, University of Zürich, 8057 Zürich, Switzerland.

CORRESPONDING AUTHOR: Damian E. Helbling, School of Civil and Environmental Engineering, Cornell University, 220 Hollister Hall, Ithaca, NY, 14853, USA. Email: damian.helbling@cornell.edu. Tel: +1 607 255 5146. Fax: +1 607 255 9004.

HIGHLIGHTS:

- Aerobic biodegradation half-lives collected from literature for eleven pesticides.
- Multivariable framework developed to link environmental metadata to half-lives.
- Application history and biomass always positively associated with half-lives.
- Relevance of other metadata depend on physicochemical properties of pesticide.
- Results provide quantitative link between half-lives and partitioning behavior.

This document is the accepted manuscript version of the following article:
Wang, Y., Lai, A., Latino, D., Fenner, K., & Helbling, D. E. (2018). Evaluating the environmental parameters that determine aerobic biodegradation half-lives of pesticides in soil with a multivariable approach. Chemosphere, 209, 430-438.
http://doi.org/10.1016/j.chemosphere.2018.06.077

ABSTRACT

Aerobic biodegradation half-lives (half-lives) are key parameters used to evaluate pesticide persistence in soil. However, half-life estimates for individual pesticides often span several orders of magnitude, reflecting the impact that various environmental or experimental parameters have on half-lives in soil. In this work, we collected literature-reported half-lives for eleven pesticides along with associated metadata describing the environmental or experimental conditions under which they were derived. We then developed a multivariable framework to discover relationships between the half-lives and associated metadata. We first compared data for the herbicide atrazine collected from 95 laboratory and 65 field studies. We discovered that atrazine application history and soil texture were the parameters that have the largest influence on the observed half-lives in both types of studies. We then extended the analysis to include ten additional pesticides with data collected exclusively from laboratory studies. We found that, when data were available, pesticide application history and biomass concentrations were always positively associated with half-lives. The relevance of other parameters varied among the pesticides, but in some cases the variability could be explained by the physicochemical properties of the pesticides. For example, we found that the relative significance of the organic carbon content of soil for determining half-lives depends on the relative solubility of the pesticide. Altogether, our analyses highlight the reciprocal influence of both environmental parameters and intrinsic physicochemical properties for determining half-lives in soil.

KEYWORDS: aerobic biodegradation; pesticide; multivariable analysis; partitioning; bioavailability

1. INTRODUCTION

Abiotic and biotic degradation are major determinants of the environmental fate of chemicals. As such, quantitative estimates of abiotic and biotic degradation half-lives or rate constants are key input parameters for environmental fate models (ECETOC, 2017). However, environmental fate models are often most sensitive to inaccuracies in degradation rate constants (Fenner et al., 2007; Horst and Koelmans, 2016) and degradation rate constants are often the most uncertain parameters included in environmental fate models (Ghafoor, 2013; Krutz et al., 2008). Much of the uncertainty resides in the fact that degradation depends on both environmental parameters and intrinsic physicochemical properties of the chemical. Therefore, degradation rate constants can range widely for a given chemical under varying environmental conditions (Aronson et al., 2006; Fenner et al., 2007). Whereas much is known about parameters that are expected to influence degradation half-lives, less is known about the quantitative effects that changes in those parameters may have (Boethling et al., 2009; Mclachlan et al., 2016).

One of the most well-studied degradation processes is the aerobic biodegradation of pesticides in soil (Pal et al., 2006). For example, atrazine biodegradation in soil has been studied for decades and hundreds of half-lives have been reported in the literature (Charnay et al., 2005; Fenner et al., 2007; Müller et al., 2003; Vischetti et al., 1997). Many of those studies likewise report associated metadata describing the environmental or experimental conditions under which the half-lives were derived. The sum of these studies confirms that aerobic biodegradation in soil is a highly variable process, with estimated half-lives ranging between one day and several years. However, the relative importance of the environmental parameters

66 that contribute to this observed variability remain poorly understood (Shaner et al., 2011) and
67 cannot be fully explained by bivariate correlation analyses (Greskowiak et al., 2017).

68 An improved understanding of pesticide persistence in soil is predicated on an improved
69 understanding of the environmental parameters that influence aerobic biodegradation half-
70 lives (half-lives). If the most important environmental parameters driving the half-life of a
71 pesticide in soil are well described, then laboratory experiments could be designed to develop a
72 quantitative understanding of those dependencies, which could in turn be incorporated into
73 environmental fate models. The primary objectives of this work were to: (i) collect half-lives
74 and associated metadata describing the environmental conditions under which they were
75 derived for a set of well-studied pesticides; and (ii) develop a multivariable framework to
76 evaluate the environmental or experimental parameters that influence the reported half-lives
77 for those pesticides. Because we collected a significant amount of aerobic biodegradation data
78 from both laboratory and field studies for atrazine, we first describe and demonstrate our
79 multivariable approach in detail by comparing the parameters that influence atrazine
80 biodegradation in laboratory and field systems. We then present the results of our
81 multivariable approach for ten more pesticides (2,4-D, diuron, metribuzin, acetochlor,
82 chlorsulfuron, mandipropamid, metamitron, metazachlor, quinmerac, and metsulfuron-methyl),
83 and discuss the environmental parameters that exhibit a significant influence on the variability
84 of degradation rate constants among all eleven pesticides.

85 **2. MATERIALS AND METHODS**

86 *2.1 Chemical selection and data collection*

We collected primary degradation half-lives or first-order rate constants representing aerobic biodegradation in soil for eleven pesticides from a variety of sources including the Eawag-Soil in enviPath database, regulatory documents, technical reports, and the scientific literature. The eleven pesticides and their physicochemical properties are listed in **Table 1**. Pesticides were selected based on data availability, but also represent a range of physicochemical properties including the octanol-water partition coefficient (K_{ow}), organic carbon-water partition coefficient (K_{oc}), and aqueous solubility. For atrazine, we collected degradation rate constants from laboratory and field studies. For the remaining pesticides, we collected degradation rate constants only from laboratory studies (due to limited availability of field data). A summary of the rate constants are provided in **Tables S1-S12** of the Supplementary Data (SD).

Associated metadata describing the environmental parameters under which the half-lives were derived were also collected. There were fourteen environmental parameters that were reported along with degradation rate constants in at least one of the datasets collected for at least one of the pesticides. These include temperature (T), pH, organic carbon content (c_{org}), sand content ($sand$), silt content ($silt$), clay content ($clay$), total nitrogen (n_{tot}), experimental moisture content as % of water holding capacity ($water$), minimum soil sampling depth which could range from 0 (surface sampling) to some positive value (d_{min}), total sampling depth when soil was collected at varying depths and mixed (d_{diff}), bulk density ($bulk$), cation exchange capacity (CEC), pesticide application history ($pest.cond$) defined as application of the pesticide on the soil within the previous four years (1) or not (0), and the biomass concentration ($biomass$). A summary of these reported data for each pesticide are provided in **Tables S1-S12** of the SD.

2.2 Modelling framework

We evaluated the data collected for the eleven pesticides by means of multivariable analyses to gain insight on the environmental parameters that influence the magnitude of the reported aerobic biodegradation rate constants. The overall modelling framework that we developed included three steps: metadata processing; multivariable regression; and confidence analysis. A schematic of the overall modelling framework is provided in **Figure 1**.

2.3 Metadata processing

Metadata processing included: (1) evaluation of covariation among each of the environmental parameters for each pesticide; and (2) imputation to generate complete datasets when the values for an environmental parameter were not reported in all datasets. We used a Pearson correlation matrix to assess covariation among the environmental parameters and selected one representative environmental parameter when significant covariation was observed. We used imputation to generate complete datasets by establishing quantitative relationships between each environmental parameter and the biodegradation rate constants (Buuren and Oudshoorn, 2011; Pampaka et al., 2016). We used multiple imputation employed with the *MICE* package in the R Statistical Software (R Core Team) using the predicted mean matching imputation method to generate five plausible values for each missing value, thus generating five complete datasets for each pesticide; the values are generated using a regression between the dependent variable (biodegradation rate constants) and the predictor variables (environmental parameters) plus or minus a random term drawn from the residuals of the regression (Allison, 2000; Hippel, 2007; Schafer, 1997). The random term accounts for uncertainty in the imputation, thus providing valid variance estimates (Buuren and Oudshoorn, 2011; Dong and Peng, 2013; Stuart et al.,

2009), and eliminates coefficient inflation that might be expected when using the dependent variable to impute predictors (Allison, 2000). If the dependent variable is not used to impute the predictor variables, then the predictor variable will be imputed as though it has no relationship with the dependent variable. Then, when the imputed data are subsequently analyzed, the estimated slope of the dependent variable on the predictor will be biased toward zero, because no dependence was assumed in the imputation (Hippel, 2007; Landerman et al., 1997).

2.4 Multivariable regression

A generalized additive model (GAM) was used to develop a multivariable relationship between our set of environmental parameters (predictor variables) and the biodegradation rate constants (response variable). Based on our expectations on how each of the predictor variables influence the magnitude of the biodegradation rate constants (Fenner et al., 2007), we developed a generic multivariable model for aerobic biodegradation in soil of the form:

$$\begin{aligned}\ln(k) = {} & k_0^{'} + f^{'}(T) + g^{'}(pH) + h^{'}\left(c_{org}\right) + b_s \ln(sand) + b_i \ln(silt) \\ & + b_c \ln(clay) + b_n \ln(n_{tot}) + b_w \ln(water) + b_m \ln(d_{min}) \\ & + b_d \ln\left(d_{diff}\right) + b_{bd} \ln(bulk) + b_{cec} \ln(CEC) \\ & + b_a(pest.cond) + b_b \ln(biomass)\end{aligned} \qquad \text{Equation 1}$$

where k is the biodegradation rate constant (calculated from half-lives assuming first order biodegradation) and $k_0^{'}$ is a model fitting constant. The terms $f^{'}(T)$, $g^{'}(pH)$, and $h^{'}\left(c_{org}\right)$ are initially unspecified functions of T, pH, and c_{org} because we expect that biodegradation rate constants will vary non-monotonically with these parameters (Fenner et al., 2007). We fit Equation 1 with the biodegradation rate constants for each pesticide and the imputed metadata using the *gam* package in R using nonparametric smoothers for the unspecified

functions for T, pH, and c_{org}. The GAM output generates plots of the shapes of the functions that best describe the contribution of each parameter to the magnitudes of the biodegradation rate constants. These plots were used to assign functions to the T, pH, and c_{org} variables and to write the final model in a closed parametric form.

2.5 Confidence analysis

The confidence analysis includes: (1) bootstrap resampling of the fully imputed datasets for each of the pesticides; and (2) stepwise linear regression to select the most significant predictor variables into multivariable models. We used bootstrap sampling with replacement to generate 10,000 individual datasets for each of the five imputed datasets (for a total of 50,000 datasets) based on the full population of data collected (or imputed) for each of the eleven pesticides. Each of the datasets included 60 values drawn from the original population regardless of the size of the original population. Replacement was applied to represent a truly random resampling, such that a later sample would not depend on the results of the initial sampling. We implemented stepwise linear regression in R to identify models that generate the highest accuracy while incorporating the fewest number of predictor variables. We applied stepwise linear regression to each of the 50,000 bootstrapped datasets for each pesticide to evaluate how frequently each of the predictor variables was selected into well-performing models. We defined well-performing stepwise regression models as those that performed as well as or better than the GAM model for each pesticide. We used the frequency that each predictor variable was selected into well-performing models as a metric of how generally important that predictor variable is in determining the magnitude of the reported biodegradation rate constants.

3. RESULTS AND DISCUSSION

3.1 General characteristics of the atrazine datasets

We first demonstrate our multivariable approach in detail by examining the environmental or experimental parameters that influence atrazine biodegradation in laboratory and field systems. If laboratory studies adequately reflect aerobic biodegradation of pesticides in the field, then we would expect that the same environmental parameters would be identified in both laboratory and field systems. We collected 95 biodegradation half-lives reported from laboratory studies and 65 biodegradation half-lives reported from field studies. Interestingly, the mean values of the half-lives are similar for laboratory and field studies (approximately 25 days), though the standard deviation of reported half-lives from laboratory studies is much larger. Specifically, laboratory studies have reported atrazine half-lives ranging between 1 to 770 days whereas field studies report half-lives between 3.5 and 277 days. It is often reported that dissipation in the field is more rapid than what is observed in the laboratory (Ismail and Kalithasan, 2006; Walker, 1987). This has been attributed to changes in soil properties during handling, changes in the microbial community, or elimination of other loss processes such as volatilization (Ismail and Kalithasan, 2006; Lay and Ilnicki, 1975). The wider range of half-lives noted in the data from laboratory studies could be the result of these or related processes.

We collected associated metadata for each of the 160 atrazine half-lives, which included at least one value for 13 of the 14 environmental parameters under consideration; only $biomass$ was unreported. Among the 13 environmental parameters, there were between 1% - 88% missing values in the laboratory studies and 18% - 91% missing values in the field studies. We compared the distributions of each predictor variable among laboratory and field studies.

For most of the predictor variables, there was no statistically significant difference in the distributions between laboratory and field studies (p>0.05, Mann-Whitney). However, we found that the distributions of T, $sand$, and d_{min} were significantly different between laboratory and field studies (p<0.05, Mann-Whitney). Those differences may further explain the wider range of half-lives reported for the data from laboratory studies.

3.2 Metadata processing

The first step in metadata processing was to evaluate covariation among the predictor variables. Pearson correlation matrices comparing the metadata show that $sand$ in laboratory and field studies significantly associates with $clay$ (ρ=-0.72 in laboratory studies and ρ=-0.69 in field studies, p<0.05) and $silt$ (ρ=-0.86 in laboratory studies and ρ=-0.87 in field studies, p<0.05). This is not unexpected based on the composition of soil and this type of association has been previously reported (Fenner et al., 2007). None of the other predictor variables were found to co-vary with any other predictor variable for which sufficient data was available. Consequently, we used $sand$ as the representative predictor for soil texture in our multivariable regression analysis.

The next step in metadata processing was to evaluate the extent of missing data for each of the remaining predictor variables and to use multiple imputation to fill in missing values for those parameters for which a sufficient amount of data was available. Whereas multiple imputation is a widely used technique to fill in missing values for incomplete datasets (Pampaka et al., 2016), there are no guidelines for determining how much data is required to achieve robust multiple imputation from an incomplete dataset (Dong and Peng, 2013). Therefore, we used the nearly complete datasets for T, c_{org}, $sand$, d_{min}, and d_{diff} from atrazine laboratory

studies to generate incomplete datasets by randomly deleting 10%, 20%, 30%, …, 90% of the values for each variable. We then used multiple imputation to fill in the missing values of each of the synthetically generated incomplete datasets and evaluated the performance of multiple imputation by calculating the correlation coefficient between the original dataset and the imputed dataset. The results of our analysis demonstrate that multiple imputation can be sufficiently accurate (R^2>0.7) when no more than 40% of the data is missing. Therefore, environmental parameters that had more than 40% missing values were excluded from further analysis. For atrazine laboratory studies, n_{tot} (83% missing), $bulk$ (92% missing) and CEC (88% missing) were excluded and for atrazine field studies, T (83% missing), n_{tot} (72% missing), $water$ (82% missing), $bulk$ (91% missing) and CEC (78% missing) were excluded from the multivariable analysis.

3.3 Multivariable regression

We fit a truncated form of Equation 1 (with parameters discarded during metadata processing removed) with the imputed datasets using the *gam* package in R. The output from the GAM allows us to examine the shape of the response of the degradation rate constants to each of the predictor variables, which is plotted in **Figure 2** for laboratory and field studies. As expected, the biodegradation rate constants varied non-monotonically with the magnitudes of T in laboratory studies and with the magnitudes of pH and c_{org} in laboratory and field studies (not enough data to evaluate T in field studies). The biodegradation rate constants from the laboratory studies reached a maximum value at a T of 25°C, similar to previous findings (Fenner et al., 2007). Under most ambient conditions, T is expected to have a positive association with degradation rates, though higher temperatures could change microbial communities and

238 negatively influence degradation rates. The non-monotonic relationship with pH was not as
239 pronounced, but a maximum value was observed at pH 6.5 in both laboratory and field studies.
240 The relationship with c_{org} is not as clear. Our expectation was that degradation rates would be
241 lower at low c_{org} due to nutrient limitation and reduced microbial activity and at high c_{org} due
242 to increased adsorption (Fenner et al., 2007). In other words, we expected to observe an
243 optimum level of c_{org} at which degradation rates would be highest. Our data show lower
244 degradation rates at high c_{org} for atrazine in laboratory and field studies, but no relationship
245 was observed at low c_{org}. Nevertheless, a non-monotonic relationship is noted between
246 biodegradation rate constants and c_{org} in both laboratory and field studies.

247 Based on the shapes of the non-monotonic relationships observed for T, pH, and c_{org}
248 we expressed these unspecified functions as second order polynomials to write the
249 multivariable model in a closed parametric form. We also factored the temperature term and
250 used T^{-1} as the predictor to facilitate comparison to the Arrhenius equation (Fenner et al., 2007).
251 The linear model suggested by GAM for atrazine laboratory studies becomes:

$$\begin{aligned}\ln(k) = {} & k_0' + b_{T1}\frac{1}{T} + b_{T2}\frac{1}{T^2} + b_{pH1}pH + b_{pH2}pH^2 + b_{c1}\ln(c_{org}) \\ & + b_{c2}\ln(c_{org})^2 + b_s\ln(sand) + b_w\ln(water) + b_m\ln(d_{min}) \\ & + b_d\ln(d_{diff}) + b_a atr.cond\end{aligned} \qquad \text{Equation 2}$$

252 The linear model suggested by GAM for atrazine field studies becomes:

$$\begin{aligned}\ln(k) = {} & k_0' + b_{pH1}pH + b_{pH2}pH^2 + b_{c1}\ln(c_{org}) + b_{c2}\ln(c_{org})^2 \\ & + b_s\ln(sand) + b_w\ln(water) + b_m\ln(d_{min}) + b_d\ln(d_{diff}) \\ & + b_a atr.cond\end{aligned} \qquad \text{Equation 3}$$

253 The linear models have R^2 values of 0.79 and 0.48 for laboratory and field studies, respectively.

254 *3.4 Confidence analysis*

255 The GAM analysis provides a general multivariable model and enables inspection of the shapes
256 of the relationships between the biodegradation rate constants and each of the predictor
257 variables, but it does not provide any information on how important each of the predictor
258 variables are in determining the magnitudes of the biodegradation rate constants. General
259 approaches to identify key predictor variables that contribute to the magnitude of a response
260 variable include stepwise linear regression or subset selection modelling, both of which have
261 recently been applied to evaluate key variables influencing biodegradation half-lives (Fenner et
262 al., 2007; Latino et al., 2017). Although these are statistically sound approaches, they produce a
263 single set of results based on the input data with no assessment of the confidence in those
264 results.

265 Bootstrap techniques add another dimension to statistical estimation strategies by
266 assessing the robustness of a particular estimator (Stine, 1989). By combining stepwise linear
267 regression with bootstrap techniques, we developed an approach that allows for a ranking of
268 predictor variables based on the frequency in which they are selected as significant parameters
269 in well-performing models. We used bootstrap resampling to generate 10,000 datasets for each
270 of the five imputed laboratory and field datasets. This generated a total of 50,000 datasets for
271 laboratory and field studies, each made up of 60 rate constants selected at random from the
272 original populations. We then used stepwise linear regression to build multivariable models
273 that had the highest accuracy (R^2) while incorporating the fewest number of predictor variables.
274 To be consistent with the results of our GAM analysis, we extracted the subset of models that
275 explained at least 80% of the variability in the rate constant data ($R^2 \geq 0.8$) for the laboratory
276 studies and 50% of the variability in the rate constant data from the field studies ($R^2 \geq 0.5$) as

well-performing models among the 50,000 models generated by stepwise linear regression. A total of 11,250 models had R^2 values greater than 0.8 for atrazine laboratory studies and 19,260 models had R^2 greater than 0.5 for atrazine field studies. Each of the stepwise linear regression models include a subset of predictor variables that are each assigned a p-value that describes how much they contribute to the performance of the multivariable model. For example, $sand$ was selected more than 10,000 times as a significant variable ($p<0.01$) among the 11,250 well-performing models, which translates to a 92% selection frequency.

Based on our integrated bootstrap and stepwise linear regression approach, the key environmental variables influencing biodegradation rate constants for atrazine in laboratory and field studies are presented in **Table 2**, where the environmental parameters are ranked from highest influence to lowest influence. The median, minimum, and maximum values of the coefficients of the parametrized models are also provided. Importantly, the data show that the environmental parameters that are relevant for atrazine biodegradation are similar in laboratory and field studies, demonstrating that the differences in rate constants measured in laboratory and field studies are likely due to the differences in the ranges of environmental parameters explored and not the result of experimental artifacts. $atz.cond$ and $sand$ were selected most frequently as significant environmental variables for both laboratory and field studies, indicating that they are the most important parameters in well-performing models under both laboratory and field conditions. Based on our finding that the ranking of each parameter under laboratory and field conditions are similar, it is reasonable to suggest that T could also be a key environmental parameter for atrazine degradation under field conditions,

but the lack of documentation of temperature data in field studies excluded it as a model input parameter.

Interestingly, the coefficients for most parameters selected into well-performing models range over positive and negative values, reflecting some sensitivity of the results to the population of data sampled during bootstrapping. However, the coefficients for $atz.cond$ and $sand$ were always positive for laboratory studies and $atz.cond$ was always positive in field studies. The changing behavior of the coefficients for $sand$ between laboratory and field studies is notable, but $sand$ is a surrogate parameter for soil texture and we interpret this result as an indication that soil texture is generally important for the aerobic biodegradation of atrazine in soil. Further, median values can adequately reflect the central tendency of the coefficients for the other parameters. In that respect, the positive values of the median coefficients for $atz.cond$ and $water$ in laboratory and field studies suggest a general positive association with these environmental parameters. Likewise, the negative values of the median coefficients for d_{diff} in laboratory and field studies suggest a general negative association with total sampling depth.

The results presented in **Table 2** could be applied to guide the selection of appropriate degradation rate constants to be used for environmental fate modelling. For example, we found that $atz.cond$, $sand$, and T were key environmental variables for predicting atrazine biodegradation rate constants. In other words, the variability among the reported half-lives for atrazine was best explained by the variability among the values of these three environmental variables. The relationships between the reported half-lives and these three environmental variables presented in **Figure 2** also match our theoretical expectations (Fenner et al., 2007;

320 Schwarzenbach et al., 2016). Based on these findings, one could use knowledge of one or more
321 of these parameters for a particular study area to select an appropriate atrazine degradation
322 rate constant for modelling atrazine fate in that study area.

323 Our results can also be interpreted to help guide the requirements of regulatory tests to
324 focus on investigating ranges of the most important environmental variables. For example,
325 instead of using only soil samples from sites that have not had atrazine application in the past
326 four years as required by current regulatory guidelines, additional soil samples with a more
327 recent history of atrazine application could be selected for testing. In that respect, laboratory
328 test guidelines should also recommend testing soil samples that cover a range of soil textures.
329 More robust data from laboratory experiments that explore these notable dependencies may
330 lead to an improved quantitative understanding that could be incorporated into environmental
331 fate models.

332 *3.5 Application of multivariable analysis to other pesticides*

333 We next applied the multivariable framework described for atrazine to 2,4-D, diuron,
334 metribuzin, acetochlor, chlorsulfuron, mandipropamid, metamitron, metazachlor, quinmerac,
335 and metsulfuron-methyl. Due to the limited availability of field data for these pesticides, we
336 applied our multivariable workflow only to data collected from laboratory studies. As described
337 for atrazine, we used one parameter to describe soil texture ($sand$ for most and $clay$ for
338 acetochlor) due to covariation and a threshold of 40% missing data for multiple imputation. A
339 statistical summary of the data available for each pesticide, the results of the GAM analysis, and
340 the range of the magnitudes of the coefficients for parameters selected into well-performing

models are provided in the SD. A summary of the frequencies in which each parameter was selected into well-performing models is provided in **Table 3**.

We first aimed to determine whether any of the observations made for atrazine could be extrapolated to the other pesticides. None of the other pesticides contained sufficient data describing $pest.cond$ to include that parameter in our multivariable models. Nevertheless, there was sufficient data on $pest.cond$ available for 2,4-D to perform a bivariate analysis with degradation rate constants. A significant positive relationship was discovered ($p<0.01$, Pearson) providing additional evidence that pesticide application history is an important variable for determining degradation rate constants. Our multivariable analysis also revealed that $biomass$ was frequently selected in well-performing models as an important parameter describing the aerobic biodegradation of mandipropamid, though there was insufficient data to include $biomass$ in other multivariable models. However, there was sufficient data on $biomass$ available for 2,4-D, acetachlor, quinmerac, and metsulfuron-methyl to perform a bivariate analysis with degradation rate constants. A positive association was discovered between reported rate constants and $biomass$ for all four additional pesticides, suggesting that biomass concentration is also a generally important variable for determining degradation rate constants.

No other parameters were found to be generally important for aerobic biodegradation of pesticides in soil. Notably, the dependencies on pH were inconsistent and not particularly strong for most pesticides (**Figures S1-S10**), which included seven neutral and four negatively charged pesticides (**Table 1**). The expected relationship between pH and half-lives is complex, with pesticide speciation influencing bioavailability and interactions with mostly negatively charged soil particles (Chaplain et al., 2011; Schwarzenbach et al., 2016) and pH also affecting

363 microbial community composition and activity (Müller et al., 2003). Because all of the pesticides
364 are neutral or negatively charged in the pH range of the soils reported, we expect that the pH
365 dependencies that are observed among these pesticides are related to microbial community
366 composition and activity. These data reflect that pH is an environmental parameter whose
367 influence on half-lives is difficult to predict.

368 We next examined the data in **Table 3** to determine whether the frequency at which any
369 parameter was selected into well-performing models associated with any of the
370 physicochemical properties of the test pesticides (**Table 1**). We found that the frequency that
371 d_{diff} was selected into well-performing models has a positive and significant relationship with
372 the K_{oc} of the pesticide ($p<0.05$) (**Figure 3a**). Further, the coefficient for d_{diff} is always positive
373 for diuron, the pesticide for which d_{diff} is most frequently selected into well-performing
374 models, but spans positive and negative values for the remaining pesticides (**Table S13**). This
375 finding demonstrates that pesticides that adsorb more strongly to soil have a greater
376 degradation potential in deeper soils. We attribute this finding to bioavailability; as soils are
377 pooled over greater depths, c_{org} of the soil mixture decreases making the pesticide more
378 bioavailable and decreases the observed half-life of the pesticide. We further confirmed this by
379 noting a consistent negative but not significant association between reported d_{diff} and c_{org}
380 values among all of our pesticide datasets.

381 We also discovered that the frequency that c_{org} was selected into well-performing
382 models has a positive and significant association with solubility among low to medium solubility
383 pesticides ($p<0.05$) (**Figure 3b**). Remarkably, for the two pairs of pesticides with nearly identical
384 solubility, the frequency in which c_{org} was selected into well-performing models was identical,

385 providing some additional validation to this observation. Whereas c_{org} is included in our
386 multivariable analysis as a second order polynomial making it difficult to interpret the
387 magnitudes of model coefficients, inspection of the shapes of the relationships between c_{org}
388 and biodegradation rate constants for each of the pesticides from the GAM output can be
389 informative. The major differences are noted in the low and high organic carbon ranges for
390 relatively low and medium solubility pesticides. The degradation rate constants for the
391 relatively soluble metribuzin, 2,4-D, and metazachlor exhibit a strong positive association with
392 c_{org} in the low organic carbon range. However, the degradation rate constants for the less
393 soluble atrazine, diuron, and acetochlor exhibit no association with c_{org} in the low organic
394 carbon range. In this low c_{org} range, the primary influence of c_{org} is expected to be positive, as
395 more organic carbon provides more nutrients to support microbial activity (Fenner et al., 2007);
396 this positive influence, however, is only observed for the more soluble pesticides in this analysis,
397 reflecting an apparent interdependency between our expectation and the solubility of the
398 pesticide. The behavior in the high organic carbon range is somewhat different. For the
399 relatively soluble metribuzin, continued increases in c_{org} result in continued increases in
400 biodegradation rate constants. The degradation rate constants for the mid-soluble pesticides
401 such as 2,4-D, acetachlor, and quinmerac exhibit no changes with c_{org} in the high organic
402 carbon range. Finally, the degradation rate constants for the least soluble pesticides such as
403 atrazine, diuron, and mandipropamid decrease with increasing c_{org} in the high organic carbon
404 range. In the high organic carbon range, the primary influence is expected to be negative as
405 more organic carbon content results in greater extents of adsorption (Lucia and Silveira, 2005;
406 Nam and Kim, 2002); we see this expectation manifest only for the less soluble pesticides while

the more soluble pesticides are unaffected by increasing organic carbon content (Jardine et al., 1989). Together these data demonstrate that a complex relationship exists between the aerobic biodegradation of pesticides in soil and c_{org} that depends on the intrinsic physicochemical properties and partitioning behavior of the pesticide. Because adsorption and bioavailability play an apparent role in these observations, we expected that a relationship would also emerge between the frequency that c_{org} was selected into well-performing models and the K_{oc} values of the pesticides. We did note a similar (but negative) association between the frequency that c_{org} was selected into well-performing models and the K_{oc} values of the pesticides, but our data suggest that solubility is the better predictor of relationships between c_{org} and half-lives, likely due to reasons of uncertainty in K_{oc} estimates (**Table 1**) and the limited number of data available (*n=8*, **Figure 3**). Nevertheless, these notable findings provide an important quantitative link between observed half-lives of pesticides in soil and our theoretical understanding of partitioning behavior and bioavailability.

ACKNOWLEDGEMENTS

This study was funded by the CEFIC Long-range initiative under project identifier LRI-ECO31. The data collection was also carried out with financial support from the European Research Council under the European Union's Seventh Framework Programme (ERC grant agreement no. 614768, PROduCTS) and the Swiss National Science Foundation (SNF project number CR23I2_140698).

426 **REFERENCES**

427 Allison, P.D., 2000. Multiple Imputation for Missing Data: A Cautionary Tale. Sociol. Methods Res. 28,
428 301–309.
429 Aronson, D., Boethling, R.S., Howard, P.H., Stiteler, W., 2006. Estimating biodegradation half-lives for
430 use in chemical screening. Chemosphere 63, 1953–1960.
431 Boethling, R.S., Fenner, K., Howard, P.H., Klecka, G., Madsen, T., Snape, J.R., Whelan, M.J., 2009.
432 Enviromental persistence of organic pollutants: guidance for development and review of POP risk
433 profiles. Integr. Environ. Assess. Manag. 5, 539–556.
434 Buuren, S. van, Oudshoorn, K., 2011. MICE: Multivariate imputation by chained equations in R. J. Stat.
435 Softw. 45, 1–67.
436 Chaplain, V., Mamy, L., Vieublé-Gonod, L., Mougin, C., Benoit, P., Barriuso, E., Nélieu, S., 2011. Chapter
437 29, Fate of Pesticides in Soils: Toward an Integrated Approach of Influential Factors, in: Stoytcheva,
438 M. (Ed.), Pesticides in the Modern World - Risks and Benefits. InTech, Croatia.
439 Charnay, M.P., Tuis, S., Coquet, Y., Barriuso, E., 2005. Spatial variability in 14C-herbicide degradation in
440 surface and subsurface soils. Pest Manag. Sci. 61, 845–855.
441 Dong, Y., Peng, C.J., 2013. Principled missing data methods for researchers. Springerplus 2, 1–17.
442 ECETOC, 2017. Biodegradation Default Half-Life Values in the Light of Environmentally Relevant
443 Biodegradation Studies, Analysis of the ECETOC biodegradation data base, Technical report No. 129.
444 Brussels, Belgium.
445 Fenner, K., Lanz, V.A., Scheringer, M., Borsuk, M.E., 2007. Relating atrazine degradation rate in soil to
446 environmental conditions: Implications for global fate modelling. Environ. Sci. Technol. 41, 2840–
447 2846.
448 Ghafoor, A., 2013. Understanding the Causes of Spatial Variation in Pesticide Sorption and Degradation
449 at the Catchment Scale. Swedish University of Agricultural Sciences.
450 Greskowiak, J., Hamann, E., Burke, V., Massmann, G., 2017. The uncertainty of biodegradation rate
451 constants of emerging organic compounds in soil and groundwater: A compilation of literature
452 values for 82 substances. Water Res. 126, 122–133.
453 Hippel, P.T., 2007. Regression with Mising Y's: An improved Strategy for Analysing Multiple Imputed
454 Data. Sociol. Methodol. 37, 265--291.
455 Horst, M.M.S. ter, Koelmans, A.A., 2016. Analyzing the Limitations and the Applicability Domain of
456 Water− Sediment Transformation Tests like OECD 308. Environ. Sci. Technol. 50, 10335–10342.
457 Ismail, B.S., Kalithasan, K., 2006. Dissipation and Mobility of Permethrin in the Field with Repeated

458 Applications Under Tropical Conditions. J. Environ. Sci. Heal. Part B B38, 133–146.
459 Jardine, P.M., Weber, N.L., McCarthy, J.F., 1989. Mechanisms of Dissolved Organic Carbon Adsorption
460 On Soil. Soil Sci. Soc. Am. J. 53, 1378–1385.
461 Krutz, L.J., Shaner, D.L., Accinelli, C., Zablotowicz, R.M., Henry, W.B., 2008. Atrazine dissipation in s-
462 triazine-adapted and nonadapted soil from Colorado and Mississippi: implications of enhanced
463 degradation on atrazine fate and transport parameters. J. Environ. Qual. 37, 848–857.
464 Landerman, L.R., Land, K.C., F., P.C., 1997. An empirical evaluation of the predictive mean matching
465 method for imputing missing values. Sociol. Methods Res. 26, 3–33.
466 Latino, D., Wicker, J., Gütlein, M., Schmid, E., Kramer, S., Fenner, K., 2017. Eawag-Soil in enviPath: a new
467 resource for exploring regulatory pesticide soil biodegradation pathways and half-life data. Environ.
468 Sci. Process Impacts 19, 449–464.
469 Lay, M.M., Ilnicki, R.D., 1975. Effect of soil storage on propanil degradation. Weed Res 15, 63–66.
470 Lucia, M., Silveira, A., 2005. Dissolved organic carbon and bioavailability of N and P as indicators of soil
471 quality. Sci. Agric. 62, 502–508.
472 Mclachlan, M.S., Zou, H., Gouin, T., 2016. Using benchmarking to strengthen the assessment of
473 persistence. Environ. Sci. Technol. 51, 4–11.
474 Müller, K., Smith, R.E., James, T.K., Holland, P.T., Rahman, A., 2003. Spatial variability of atrazine
475 dissipation in an allophanic soil. Pest Manag. Sci. 59, 893–903.
476 Nam, K., Kim, J.Y., 2002. Persistence and bioavailability of hydrophobic organic compounds in the
477 environment. Geosci. J. 6, 13–21.
478 Pal, R., Chakrabarti, K., Chakraborty, A., Chowdhury, A., 2006. Degradation and Effects of Pesticides on
479 Soil Microbiological Parameters-A Review. Int. J. Agric. Res. 1, 240–258.
480 Pampaka, M., Hutcheson, G., Williams, J., 2016. Handling missing data : analysis of a challenging data set
481 using multiple imputation. Int. J. Res. Method Educ. 39, 19–37.
482 Schafer, J., 1997. Analysis of incomplete multivariate data. Chapman and Hall, London.
483 Schwarzenbach, R., Gschwend, P., Imboden, D., 2016. Environmental organic chemistry, Third. ed. John
484 Wike&Sons Ltd, New Jersey.
485 Shaner, D.L., Stromberger, M., Khosla, R., Helm, A., Bosley, B., Hansen, N., 2011. Spatial distribution of
486 enhanced atrazine degradation across northeastern Colorado cropping systems. J. Environ. Qual.
487 40, 46–56.
488 Stine, R., 1989. An introduction to bootstrap methods, examples and ideas. Sociol. Methods Res. 18,
489 243–291.

490 Stuart, E.A., Azur, M., Frangakis, C., Leaf, P., 2009. Practice of Epidemiology Multiple Imputation With
491 Large Data Sets : A Case Study of the Children ' s Mental Health Initiative. Pract. Epidemiol. 169,
492 1133–1139.
493 Vischetti, C., Businelli, M., Marini, M., 1997. Characterization of spatial variability structure in three
494 separate field trials on pesticide dissipation. Pestic. Sci. 50, 175–182.
495 Walker, A., 1987. Evaluation of a simulation model for prediction of herbicide movement and
496 persistence in soil. Weed Res. 27, 143–152.
497

FIGURE CAPTIONS

Figure 1. A schematic of the overall modelling framework. Degradation data and associated metadata are collected from laboratory studies (all pesticides) or field studies (atrazine only). Pearson correlations are utilized to test for covariance among the metadata parameters (predictor variables). Multiple imputation is used to fill-in missing data values for predictor variables with less than 40% missing data and generate five complete datasets. GAM parameterizes a multivariable linear model and provides shape functions that describe the relationship between the magnitudes of degradation rate constants and each of the predictor variables. Bootstrap resampling generates 10,000 datasets from each of the five fully imputed datasets. Stepwise linear regression constructs multivariable linear models that contain the fewest number of predictor variables while meeting defined performance criteria. The predictor variables are ranked by the frequency in which they are selected into well-performing models, providing a level of confidence for identifying the environmental parameters that influence the magnitude of the reported aerobic biodegradation rate constants.

Figure 2. Shape functions describing the relationships between biodegradation rate constants and each of the predictor variables included in GAM models for atrazine in laboratory studies (left) and field studies (right). The vertical axes indicate the contribution of each variable to the value of ln(k). Points represent experimental values and partial residuals, solid lines represent the model fit, and dotted lines represent confidence intervals of ±1 standard error.

Figure 3. Relationships between (a) the percentage of well-performing models that select d_{diff} as a predictor variable and the K_{oc} of the pesticides and (b) the percentage of well-performing models that select c_{org} as a predictor variable and the solubility of the pesticide.

Table 1. The eleven pesticides and their physicochemical properties.

Compound	pK_a[a]	Charge at pH=7	Log K_{ow}[b]	K_{oc}[c] (mL/g)	Solubility[b] (mg/L)	Half-life[d] (days)
Atrazine	2.7	neutral	2.6	39-155	35	72
2,4-D	2.8	negative	2.8	20-100	677	12.4
Diuron	13	neutral	2.7	480	42	302
Metribuzin	2.5	neutral	1.7	60	1050	14.5
Acetochlor	15	neutral	3.0	139	223	11.5
Chlorsulfuron	2.5	negative	2.0	45-110	28000	66.9
Mandipropamid	15	neutral	3.6	405-1294	2	65.7
Metamitron	2.8	neutral	0.8	23-133	1800	24.9
Metazachlor	2.3	neutral	2.1	80	430	19.9
Quinmerac	3.5	negative	2.9	19-185	223	35.4
Metsulfuron-methyl	3.5	negative	2.2	4-345	9500	27.3

[a]Data are estimated values from MarvinSketch 17.2.20 by chemaxon (http://www.chemaxon.com);
[b]Data are experimental values reported in the United States Environmental Protection Agency's EPISuite software;
[c]Data are from Open Chemistry Database, National Institutes of Health, National Centre for Biotechnology Information. We used the midpoint of the given range for association testing;
[d]Data are average values of the collected datasets in this study.

Table 2. Summary of the frequencies in which each parameter was selected into well-performing models for atrazine laboratory and field data.

	Atrazine Laboratory Studies		Atrazine Field Studies	
Environmental Variables[a]	**Selection Frequency[b]**	**Coefficients (median, minimum, maximum)[c]**	**Selection Frequency**	**Coefficients (median, minimum, maximum)**
atz.cond	99%	2.3 1.2 3.8	99%	1.4 0.5 2.5
sand	92%	1.2 0.3 4.1	57%	-0.7 -2.0 0.5
1/T	76%	-36 -340 430	NA	NA
$1/T^2$	45%	130 -4500 2900	NA	NA
d_{min}	45%	-0.2 -0.4 0.3	0%	NA
pH	34%	4.1 -6.3 19	9%	2.4 -14 11
pH^2	29%	-0.3 -1.4 0.4	11%	-0.2 -0.9 1.0
C_{org}	23%	0.6 -0.6 1.3	21%	0.06 -0.7 0.9
C_{org}^2	12%	0.2 -0.5 1.0	30%	0.03 -0.1 0.2
water	20%	0.3 -0.4 0.8	NA	NA
d_{diff}	7%	-0.3 -0.9 0.6	7%	-0.4 -1.9 0.9

[a1]Environemntal or experimental parameters included in multivariable analysis;

[b]Frequency at which each environmental or experimental parameter was selected into well-performing models at a significance level of <0.01;

[c]Range of coefficients among well-performing models; NA indicates the parameter was not included in multivariable analysis due to missing values.

Table 3. Summary of the frequencies in which each parameter was selected into well-performing models for ten additional pesticides.

	2,4-D	**Diuron**	**Metri-buzin**	**Aceto-chlor**	**Chlor-sulfuron**	**Mandi-propamid**	**Meta-mitron**	**Metaza-chlor**	**Quin-merac**	**Met-sulfuron-methyl**
Observations, n	74	53	36	49	51	23	28	49	27	14
R^2 of GAM	0.27	0.74	0.65	0.23	0.78	0.78	0.66	0.72	0.82	0.88
% models > R^2 of GAM[a]	82%	42%	70%	60%	62%	88%	99%	59%	84%	35%
pest.cond	NA[b]	NA	NA	NA	NA	NA	NA	NA	NA	NA
sand[c]	9%	NA	13%	24%	NA	55%	33%	55%	53%	52%
1/T	10%	18%	36%	15%	99%	41%	66%	98%	96%	0%
$1/T^2$	10%	29%	49%	14%	97%	34%	41%	98%	87%	0%
d_{min}	NA	0%	26%	NA	NA	NA	NA	0%	NA	NA
pH	15%	3%	24%	13%	41%	76%	31%	15%	29%	51%
pH^2	19%	5%	16%	21%	73%	77%	36%	9%	47%	71%
c_{org}	74%	23%	66%	40%	19%	46%	14%	39%	41%	NA
c_{org}^2	88%	28%	20%	2%	24%	56%	12%	72%	73%	NA
water	36%	NA	NA	58%	9%	7%	69%	31%	NA	48%
d_{diff}	NA	99%	14%	16%	NA	NA	NA	43%	NA	NA
CEC	NA	NA	NA	NA	NA	4%	NA	NA	NA	NA
bulk	NA	NA	NA	NA	NA	0%	NA	NA	NA	NA
biomass	NA	NA	NA	NA	NA	2%	NA	NA	NA	NA

[a]% models > R^2 of GAM refers to the percentage of models that had an R^2 greater than the GAM model following confidence analysis; only these models were considered for further analysis.

[b]Frequency at which each environmental or experimental parameter was selected into well-performing models at a significance level of <0.01; NA indicates the parameter was not included in multivariable analysis due to missing values.

[c]clay used as soil texture parameter based on data availability and covariation.