Predictive quality of 26 pesticide risk indicators and one flow model: A multisite assessment for water contamination

Frédéric Pierlot, Jonathan Marks-Perreau, Benoit Real, Nadia Carluer, Thibaut Constant, Abdeljalil Lioeddine, Paul van Dijk, Jean Villerd, Olivier Keichinger, Richard Cherrier, et al.

To cite this version:

HAL Id: hal-01652920
https://hal.archives-ouvertes.fr/hal-01652920
Submitted on 30 Nov 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Predictive quality of 26 pesticide risk indicators and one flow model: A multisite assessment for water contamination

Frédéric Pierlot a,b,⁎, Jonathan Marks-Perreau c, Benoît Réal c, Nadia Carluer d, Thibaut Constant e, Abdeljalil Lioeddine e, Paul van Dijk f, Jean Villerda a, Olivier Keichinger g, Richard Cherrier b, Christian Bockstaller h

a LAE, Université de Lorraine, INRA, 54500, Vandoeuvre, France
b Chambre Régionale d’Agriculture Grand Est, Pôle Recherche Développement et Innovations, France
c Arvalis - Institut du Végétal, France
d IRSTEIAquatic Ecosystems, Ecology and Pollution Lyon, France
e In Vivo AgroSolutions, France
f Association pour la Relance Agronomique en Alsace, France
g Independent Researcher, France
h LAE, INRA, Université de Lorraine, 68000 Colmar, France

HIGHLIGHTS
• Information on predictive quality of pesticide risk indicators is scarce
• Outputs of 26 indicators and 1 model were compared to pesticide measurements in water
• 3 comparison tests were performed for a dataset of 1040 measurements from 3 sites
• Predictive quality was low to medium for the indicators and acceptable for the model
• The model and indicators with medium predictive quality can be recommended for use

ABSTRACT
Stakeholders need operational tools to assess crop protection strategies in regard to environmental impact. The need to assess and report on the impacts of pesticide use on the environment has led to the development of numerous indicators. However, only a few studies have addressed the predictive quality of these indicators. This is mainly due to the limited number of datasets adapted to the comparison of indicator outputs with pesticide measurement. To our knowledge, evaluation of the predictive quality of pesticide indicators in comparison to the quality of water as presented in this article is unprecedented in terms of the number of tested indicators (26 indicators and the MACRO model) and in terms of the size of datasets used (data collected for 4 transfer pathways, 20 active ingredients (a.i.) for a total of 1040 comparison points). Results obtained on a.i. measurements were compared to the indicator outputs, measured by: (i) correlation tests to identify linear relationship, (ii) probability tests comparing measurements with indicator outputs, both classified in 5 classes, and assessing the probability i.e. the percentage of correct estimation and overestimation (iii) by ROC tests estimating the predictive ability against a given threshold. Results showed that the correlation between indicator outputs and the observed transfers are low ($r < 0.58$). Overall, more complex indicators taking into account the soil, the climatic and the
1. Introduction

Ever since the end of the Second World War, widespread use of pesticides is one factor that has led to an incredible rise and securing of agricultural yields. Nevertheless, side effects on the environment (Richardson 1998) and in particular on water quality (Flury et al., 1995; Real et al., 2005; Grung et al., 2015; Lopez et al., 2015) have been observed. Consequently, regulations have been strongly reinforced, first by the European Water Framework Directive 2000/60/CE followed by various action plans which have come into existence, such as the Pesticide Package 2009/128/CE. In all cases, stakeholders involved in actions to reduce the use and impact of pesticides need operational tools to assess crop protection strategies in regard to environmental impact. The aim of such assessment may be to monitor and to report on the current status of water bodies quality, to produce references for the good management of crop protection and to work on innovative systems (Bockstaller et al., 2015).

The need of assessment tools for the pesticide issues has led to the development of numerous indicators. The simplest ones rely on and take into account the amount of quantities supplied, the Quantity of Active Ingredients (QAI) or the Treatment Frequency Index (TFI) calculating the ratio of applied pesticide to the registered rate. Although these indicators have been developed to describe the evolution of pesticide use intensity, they are often used as main indicators to address the environmental effects due to pesticide spraying in environmental assessment method (Eckert et al., 2000; Vilain et al., 2008). Pesticide risk indicators (Levitan, 2000) addressing complementary variables such as active ingredient properties, crop management data and pedoclimatic variables are more elaborate and were reviewed by several authors (Maud et al., 2001; Reus et al., 2002; Feola et al., 2011; Keichinger et al., 2013). However, these reviews have remained mainly descriptive, without providing thorough assessment of the strengths and weaknesses of the indicators. An important point in such assessment is to deal with the predictive quality of the indicators as recommended by (Bockstaller et al., 2008). Such studies have been conducted to assess the predictive quality of dynamic transfer models but on a relatively small number of active ingredients (Vanclooster et al., 2000). Stenrod et al. (2008) compared 2 indicators (EIQ and NRI) and 1 model (SWAT) to the measurement of pesticide concentration at the outlet of two watersheds. However, only one active ingredient (MCPA) was monitored. In the absence of measured data, outputs of indicators were compared between them (Maud et al., 2001; Reus et al., 2002; Feola et al., 2011), as recommended by Bockstaller and Girardin (2003). The paucity of references is therefore explained by the lack of measure datasets adapted to pesticide measurement in water with the comparison of indicator outputs.

Here we present a study aiming to assess the predictive quality of a set of pesticide risk indicators partly taken from the reviews of Devillers et al. (2005) and Keichinger et al. (2013), both based on international literature. We tried to cover the whole gradient of complexity of existing indicators. To extend this gradient, we added to our study one of the most frequently implemented models, the physically based one dimensional simulation model of vertical water and pesticide flow MACRO (Larsbo et al., 2005). The data set used for the comparison was to our knowledge unprecedented in terms of size and diversity (number of active ingredients (a.i.), pedoclimatic contexts and transfer pathways).

2. Materials and methods

2.1. Measurement of water contamination

Data from 3 different sampling sites were available, namely: La Jaillière, where pesticide transfers by drainage and runoff (mainly by saturation) are monitored since 1994; Le Magneraud where measurements of pesticide transfers by percolation have been performed since 2001; and Geispitzen, where transfers by hortonien runoff were monitored between 2000 and 2012 (Fig. 1). La Jaillière and Le Magneraud sites are managed by the cereals growers’ technical institute, Arvalis - Institut du Végétal, while the Geispitzen site was managed in collaboration between Arvalis - Institut du Végétal and a regional association, the Association pour la Relance Agronomique en Alsace (ARAA).

As shown on Fig. 1 and described below, these sites cover different soil and climatic contexts of France and different transfer pathways of pesticide to water bodies (surface water and groundwater). The outcomes are considered over a period of no more than one year after the date of application. For each application of a.i., the monitoring was stopped when the a.i. was not detected for 4 consecutive weeks. The data set is collected on a weekly basis for the Jaillière and the Magneraud sites, and according to the runoff events on the Geispitzen site. During the monitoring period, pesticide measurements were performed by an external certified laboratory that provided detection thresholds evolving from 0.05 μg/L to 0.01 μg/L or 0.02 μg/L depending on the active ingredient (except for the glyphosate and its degradation product AMPA which both have a threshold of 0.1 μg/L). The calculation of the following variables further referred to as “measured variables” was carried out from the pesticide measurements during the monitoring period for each a.i. on each plot:

i) frequency of exceedance of the threshold of the water quality standard of drinking water: 0.1 μg/L (fd1)

\[ \text{fd1} = \frac{n_{1ijk}}{n_{ijk}} \]  

(Eq 1)

with \( n_{1ijk} \): number of measurements with concentration > 0.1 μg/L for active ingredient i on plot j at sampling time k; \( n_{ijk} \): total number of measurements for active ingredient i on plot j and sampling time k

ii) maximum concentration of active ingredient measured in μg/L (cmax)

\[ \text{cmax} = \text{MAX}(c_{ijk}) \]  

(Eq 2)

with \( c_{ijk} \): concentration of active ingredient i on plot j and sampling time k

iii) maximum flux measured in mg/ha (fmax)

\[ \text{fmax} = \text{MAX}(f_{ijk}) \]  

(Eq 3)

with \( f_{ijk} \): flux of active ingredient i on plot j and sampling time k; \( f_{ijk} = c_{ijk} \cdot W_{jk} \) with \( W_{jk} \): water flux (drainage or runoff) from plot j during sampling time k.

© 2016 Elsevier B.V. All rights reserved.
iv) cumulated flux of active ingredient in mg/ha \((f_{\text{total}})\) during the measurement period

\[
f_{\text{total}} = \sum f_{ijk}
\]

(Eq. 4)

v) weighted average concentration on the period in \(\mu\text{g}/\text{L}\) \((\text{cmp})\)

\[
\text{cmp} = \frac{\sum c_{ijk} \cdot w_{jk}}{\sum w_{jk}}
\]

(Eq. 5)

The measurements were performed at field level for the sites of La Jaillière and Geispitzen and for lysimeter of 1 m² for le Magneraud site because the majority of the studied pesticide risk indicators were designed to be used at those scales. This is also the basic level for pesticide management by farmers. As it came out from the reviews previously mentioned, very few indicators address the level of watershed due to the complexity of transfer processes (Wohlfahrt et al., 2010), although this scale is relevant for water quality assessment. But this is out of the scope of this article.

2.1.1. The Jaillière site

The Jaillière experimental station is located in the Loire-Atlantique region and is under the influence of oceanic climate. Average annual rainfall is 734 mm. The Jaillière site is composed of brown hydromorphic clay-textured soil, resulting from alterite shale. This experimentation site consists of 10 agricultural plots of 0.5 to 1 ha each, where drainage and runoff by saturation waters are collected separately. During the monitoring period, the crop rotation system was a sequence of maize, winter wheat and spring or winter pea. The main objective of this site is to quantify the loss of minerals and pesticides in the water leaving the plots, and then relate this quantity with the implemented farming practices. Drainage and runoff water are routed by sealed collectors from the parcels all the way to flow measurement workrooms, which allow continuous monitoring of the evolution of the water flow along the agricultural farmsteads. Water samples are automatically collected depending on the water flow. Then weekly based water samples are sent to a laboratory for analysis. On this site, the drainage and runoff water flows to join the hydrographic network (Marks Perreau et al., 2013). The site database consisted of 273 (drainage) and 230 (runoff) applications and transfer measurements for 18 active ingredients used between 1993 and 2010 (Table 1).

2.1.2. The Magneraud site

The experimental station of the Magneraud is also under the influence of oceanic climate. Average per annum rainfall is 822 mm. The gravelly Poitou Charentes region is composed mainly of clayey and silty limestone soil developed on sand-stone strata characterized by alternating layers of hard limestone and marl. This site is made up of 14 “open” lysimetric plots of 1 m² surface, with no vertical walls and no soil shuffle. The boxes are integrated in 60 m² plots which are cultivated with help of farm equipment thus enabling measurement of water flow. During the monitoring, the crop rotation system was a sequence of maize, winter wheat and, occasionally, peas. The water which seeps down the meter thick soil column is collected in graduated cylinders. The quantity of percolated water was measured once or twice a week during the heavy-rainy season. Once the water quantities have been measured, a liter of water was sampled and sent to the laboratory for analysis. On this site, the seeped water joins groundwater located at a depth of approximately 15 m. The database used consisted of 467 pesticide applications and transfer measurements collected between 2001 and 2010 corresponding for 15 active ingredients used (Table 1).
2.1.3. The Geispitzen site

The Geispitzen experimental station is located in the hills of the lower Sundgau district (Alsace region) and has an attenuated oceanic climate, with an average annual rainfall of 770 mm. The precipitation is generally low in winter and high in late spring and summer. The climate, with an average annual rainfall of 770 mm. The precipitation lower Sundgau district (Alsace region) and has an attenuated oceanic climate, with an average annual rainfall of 770 mm. The precipitation

2.2. Selection of indicators

The indicators tested in this study cover the whole gradient of complexity of existing indicators (Bockstaller et al., 2015) which are available, selectable to researches, farmer’s advisers and water managers. The set of selected indicators include: i) indicators based on management data (TI, QA1) using only data of a.i. amount; ii) indicators combining management data and a.i. properties (SIRIS, EIQ, ADSCOR) or based on transfer coefficient (NRI); iii) qualitative predictive indicators assessing separately by decision trees the effect of soil and climate data on the one hand, and the effect of management and a.i. properties on the other hand before aggregating them (DAEG, ARTHUR); iv) qualitative predictive indicators integrating directly by fuzzy decision tree all type of variables designed by experts such as I-PHY1, I-PHY2s (Bockstaller et al., 2008), or derived by supervised learning from mechanistic models such as I-PHY2v; v) indicators derived from simplified quantitative models such as POCER or more complex such as SYNOPSIS (Gutsche and Rossberg, 1997) and EPRIP, or metamodel from a mechanistic model such as DRAINAGE HAIR (see Fig. 2). Input variables required for calculation of each indicator are described in Table A1 (Supplementary Materials) and, the equation or literature resource is given for each indicator or sub-indicator in Table A2 (Supplementary Materials). It should be noticed that rainfall data for the whole year or shorter period (e.g. 15 days), or maximum daily rainfall were used by indicators of the groups iii) and v), I-PHY2v and DRAINAGE HAIR were designed from simulations of the MACRO model. While the mechanistic model itself requires weather data for the calculation of the transfers, the derived indicators do no need anyone for their calculation. Thus, we also integrated the model MACRO (Larsho et al., 2005) in order to complete the complexity gradient.

2.3. Parameterization of indicators

2.3.1. Rules of calculation

Some of the selected indicators chosen such as I-Phy assess the risk represented by pesticides, in different environmental compartments - for example surface water, groundwater or air - and target organisms - for example human beings or aquatic organisms -. In this case, and in accordance with recommendations made by Bockstaller and Girardin (2003), we considered for the studied indicators only the sub-indicators addressing transfer of active ingredients to water. For the DEXIPM indicator based on a decision tree (Pelzer et al., 2012), only the branches of the tree dealing with pesticides and the water quality were considered. However, it was not possible to separate the transfer model from the toxicity variable, so that we calculated DEXIPM indicator at 2 toxicity levels, LOW and HIGH. For the La Jaillière (drainage) and Le Magneraud sites, only the sub-indicators estimating vertical transfer by leaching were calculated. Although pesticide transfers by drainage impact surface water, the tested indicators consider this kind of transfer with the sub-indicators for vertical transfer. For the La Jaillière (runoff by saturation of the profile) and Geispitzen (hortonian runoff), the sub-indicators considering surface transfer were calculated. In many cases, no tools to calculate the chosen sub-indicators were available, so that to optimize this step, we developed a computing chain. The whole set of variables needed to calculate selected indicators were compiled in a unique Excel® spreadsheet that was linked to single spreadsheets calculating the indicator. We checked the relevance of the calculation outputs by comparing them to the outputs from the original software (e.g., I-PHY1 and 2) or with examples of

Table 1
List of active ingredients studied: properties and application period.

<table>
<thead>
<tr>
<th>Active ingredient</th>
<th>Dosage (g/ha)</th>
<th>Koc (mL/g)</th>
<th>Field DT50 (days)</th>
<th>Application period</th>
<th>Crop types</th>
<th>Site of study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acetonifin (H)</td>
<td>300–2400</td>
<td>7126</td>
<td>80.4</td>
<td>Spring/fall/winter</td>
<td>M/SP/WP/F</td>
<td>La Jaillière/Le Magneraud</td>
</tr>
<tr>
<td>Alachlor (H)</td>
<td>2160–2328</td>
<td>124</td>
<td>14</td>
<td>Spring</td>
<td>M</td>
<td>Geispitzen</td>
</tr>
<tr>
<td>Atrazine (H)</td>
<td>250–750</td>
<td>100</td>
<td>29</td>
<td>Spring</td>
<td>M</td>
<td>Geispitzen</td>
</tr>
<tr>
<td>Bentazon (H)</td>
<td>261–1740</td>
<td>515</td>
<td>10</td>
<td>Spring/winter</td>
<td>M/SP/WP</td>
<td>La Jaillière/Le Magneraud</td>
</tr>
<tr>
<td>Bromoxynil (H)</td>
<td>62–400</td>
<td>173.5</td>
<td>5.56</td>
<td>Spring/fall/winter</td>
<td>WW/M</td>
<td>La Jaillière/Le Magneraud/Geispitzen</td>
</tr>
<tr>
<td>Chlorothalonil (F)</td>
<td>375–2250</td>
<td>850</td>
<td>44</td>
<td>Spring/winter</td>
<td>WW/SP/WP</td>
<td>La Jaillière/Le Magneraud</td>
</tr>
<tr>
<td>Diflufenicanil (H)</td>
<td>20–187</td>
<td>3416</td>
<td>415</td>
<td>Fall/winter</td>
<td>WW</td>
<td>La Jaillière/Le Magneraud</td>
</tr>
<tr>
<td>Dm-Ap (H)</td>
<td>720–1008</td>
<td>227</td>
<td>7</td>
<td>Spring</td>
<td>M</td>
<td>La Jaillière/Le Magneraud</td>
</tr>
<tr>
<td>Epoxiconazol (F)</td>
<td>25–87</td>
<td>1073</td>
<td>116.8</td>
<td>Spring/winter</td>
<td>WW</td>
<td>La Jaillière/Le Magneraud</td>
</tr>
<tr>
<td>Glysophosate (H)</td>
<td>480–1080</td>
<td>21,699.44</td>
<td>31.5</td>
<td>Spring/fall/winter</td>
<td>WW/M/SP/CIPAN</td>
<td>La Jaillière</td>
</tr>
<tr>
<td>Isoproturon (H)</td>
<td>500–1500</td>
<td>122</td>
<td>22.5</td>
<td>Fall/winter</td>
<td>WW</td>
<td>La Jaillière/Le Magneraud</td>
</tr>
<tr>
<td>Methotrinone (H)</td>
<td>30–150</td>
<td>109</td>
<td>5</td>
<td>Spring</td>
<td>M</td>
<td>La Jaillière/Le Magneraud/Geispitzen</td>
</tr>
<tr>
<td>S-metolachlor (H)</td>
<td>983–1646</td>
<td>200</td>
<td>21</td>
<td>Spring</td>
<td>M</td>
<td>La Jaillière/Le Magneraud/Geispitzen</td>
</tr>
<tr>
<td>Methylsulfuron-methyl (H)</td>
<td>5–30</td>
<td>39.5</td>
<td>31.97</td>
<td>Fall/winter</td>
<td>WW</td>
<td>La Jaillière/Le Magneraud</td>
</tr>
<tr>
<td>Nicosulfuron (H)</td>
<td>20–30</td>
<td>20.7</td>
<td>19.3</td>
<td>Spring</td>
<td>M</td>
<td>La Jaillière/Le Magneraud/Geispitzen</td>
</tr>
<tr>
<td>Pendimethalin (H)</td>
<td>250–400</td>
<td>15,744</td>
<td>99.17</td>
<td>Spring/fall/winter</td>
<td>WW/SP/WP</td>
<td>La Jaillière/Le Magneraud</td>
</tr>
<tr>
<td>Prochloraz (F)</td>
<td>315–450</td>
<td>2225</td>
<td>345.5</td>
<td>Spring/winter</td>
<td>WW</td>
<td>La Jaillière/Le Magneraud</td>
</tr>
<tr>
<td>Prosulfocarb (H)</td>
<td>800–3200</td>
<td>1693</td>
<td>9.8</td>
<td>Fall/winter</td>
<td>WW</td>
<td>La Jaillière</td>
</tr>
<tr>
<td>Prosulfuron (H)</td>
<td>3.6–15</td>
<td>16.67</td>
<td>16.44</td>
<td>Spring</td>
<td>M</td>
<td>La Jaillière/Le Magneraud/Geispitzen</td>
</tr>
<tr>
<td>Tau-fluvalinate (I)</td>
<td>48–72</td>
<td>504,123</td>
<td>90.8</td>
<td>Spring/fall</td>
<td>WW/SP</td>
<td>La Jaillière</td>
</tr>
</tbody>
</table>

(H): Herbicide/ (I): Insecticide/ (F): Fungicide.
developed in the HAIR project from the MACRO model (Strassemeyer and Gutsche 2010). For the other indicators, the description is available on the GUIDE tool (Keichinger et al. 2013): http://www.plage-evaluation.fr/guide/). Key:*: multiplication, [ ]: Decision Tree, [ ] X-[ ]: separation in 2 sub-indicators, LF: fuzzy logic, f ( ): function, * sub-indicator for vertical transfer to toxicity for DEXiPM.

(calibration. The outputs of the model used for comparison are fd1, tation situation with data available in a national database, without any we assessed the predictive quality of the model in a routine implemen-

in the soil and weather databases of Arvalis - Institut du Végétal. Thus

calculations to verify the consistency of the results. These were designed using the description of the calculation method of each indicator, found in their original publications and in the factsheets of the GUIDE tool (http://www.plage-evaluation.fr/guide/ Keichinger et al., 2013). When the information needed to calculate an indicator was unavailable or unprecise, we had to set the value a few input variables: for indicators ARTHUR and EPRIP 2 indicators, distance to the next water body was reduced to 0 m because the water fluxes were collected directly at the field outlet. The time between the date of application and the transfer fluxes was set at 3 days for EPRIP 2 (Trevisan et al., 2009) and I-PHY 2 (Wohlfart, 2008), a worst case situation that was also assumed by Strassemeyer et al. (2003).This may lead to an overestimation of the risk of the transfer by runoff for these indicators, which is preferable to underestimation (see Section 2.4.2). In a few cases, we exchanged di-

the information needed to calculate an indicator was unavailable or

Table A3)). The input soil variables needed to run the model were:
i) for each soil layer, depth, texture, stoniness, pH, organic matter content, ii) bedrock nature, and daily weather data, rain, evapotranspiration, minimum and maximum temperature. These data were collected in the soil and weather databases of Arvalis - Institut du Végétal. Thus we assessed the predictive quality of the model in a routine implementa-

First we estimated the variability of the measured data and the vari-

2.3.2. Selection of active ingredients

The active ingredients selected for our analysis on the 3 monitored sites were mainly herbicides, with few fungicides and an insecticide, be-

cause herbicides are the main source of contamination of water bodies (Lopez et al., 2015). As shown in Table 1, they cover a wide range of the main physicochemical properties involved in transfer, DT50, Koc, and solubility (Chen et al., 2015), and are applied during autumn-

winter and/or spring on different crop types and at different rates. Thus we covered a wide range of condition to get sufficient variability for the indicator results. This indicates that we have a good representation of the way the transfer of pesticides reacts in a variety of conditions and at the different periods when they are applied.

Active ingredient properties were taken from the SIRIS French data-

base, also used for an indicator developed with the same name (Gouzy and Le Gall, 2007). We tested the effect of the choice of value for Koc and DT50 settings on the correlations between measured variables and indi-

ator output for more favorable values, mainly taken from the Footprint database (PPDB: Pesticide Properties Database 2015: http:// sitem.herts.ac.uk/aeru/ppdb/en/index.htm) and more unfavorable values, mainly taken from the database of the I-PHY indicator (Bockstaller et al., 2008). Small differences (variations of correlation coefficients under|0.1|) were found except for I-PHY1 indicator at the La Jaillière site (drainage) for the favorable scenario (decrease of the r co-

efficient of 0.22), taking results obtained with SIRIS values as a refer-

ence. This was mainly due to the high weight on the results of the GUS variable aggregating Koc and DT50 (van der Werf and Zimmer, 1998). Thus, in the rest of article, the results obtained with data from the SIRIS database will be shown.

2.4. Comparison method

First we estimated the variability of the measured data and the vari-

ability of the indicator or model results in order to verify whether the ranges of variation were enough to perform the comparison. This was the case with a majority of variation coefficients higher than 50%. (Sup-

plementary Materials: Table A4). Then we performed several tests com-

paring measured variables with indicator outputs, described previously with indicators or MACRO outputs. These tests of decreasing power follow-
were: (i) correlation tests to identify linear relationships, (ii) probability tests to compare measurements with classified indicator outputs, to assess the proportion of acceptable cases (Bockstaller and Girardin, 2003) (iii) Receiver Operating Characteristic tests (ROC tests) estimating the predictive ability against a given threshold as proposed by Makowski et al. (2009). This gradient in the power of the tests aimed to identify looser relationships than linear correlations for simplified indicators.

2.4.1. Correlation test
In a preliminary step, we assessed pairwise the correlation between the indicator outputs and i) outputs of the other indicators to identify indicators showing similar results, and ii) measured variables, using successively the Pearson Correlation test and the Spearman correlation tests on the ranks. The latter, used by Reus et al. (2002) to compare outputs of pesticide risk indicators, yielded lower correlations than the Pearson test (results not shown). For the La Jaillière and Le Magnéraud sites, the size data sets allowed to calculate average value of the indicators and the measured variables for each a.i. across the years. These averages were obtained for a number of data varying between 2 and 58 (Supplementary Materials: Table A5). The Pearson Correlation test was performed for these values. For all tests we used the correlation coefficient (r) to assess the degree of relationship and not the determination coefficient (r²) that characterizes to which extent one variable explains the other. All calculations were run by means of the R software (3.2.2 version https://www.r-project.org). The result not in italic in the tables are significant by having a p value < 0.01.

2.4.2. Probability test
Bockstaller et al. (2008) define the test as assessment of the proportion of cases in which the difference predicted value – observed value falls within a probability or acceptance area defined in function of the indicator outputs and expected performance of the indicator. Since most of indicator outputs were qualitative, and following Brown et al. (2002), we ran the comparison with classes of measurements and indicator output. Like Pervanchon et al. (2005) we worked with five classes to identify a possible qualitative relationship between transfer measurements and indicator outputs. Measured variables were divided into five classes that were equivalent for fdi1: each class corresponding to a range of 20%. For the other variables, the five classes were organized following a logarithmic scale of 10 order (Table 2). Such a range of variation was also used by Brown et al. (2002) to assess their prediction model of pesticide transfer at catchment scale. Indicators outputs in form of scores were split into equal classes while indicators providing quantitative such as in EPRIP, SYNOPS and Drainage HAIR were classified like the corresponding measured variable (see Table 2).

In this study we decided that the acceptance area was defined by the sum of the correct estimations and overestimations of transfer by an indicator (Supplementary Materials: Table A6). Our choice was guided by the loose relationships between indicator’s issues and transfer of pesticides due to simplification in the design of the indicators: for example most of the indicators do not take into account the period of application. As a consequence, they do not take into account the variability of degradation processes impacted by different temperatures or different pluvimetry. The probability will be defined hereafter as the sum of correct estimation and overestimation. Discussions with water managers from the Eau Rhin-Meuse agency confirmed that overestimation of risks could be tolerated, but not underestimation. We decided to consider as acceptable indicators with a probability higher than 60% and with a percentage of correct estimation higher than 50%. These thresholds should not be considered as absolute. They allowed us to consider as acceptable only indicators which had a majority of correct estimation and to limit the degree of underestimation. This excludes indicators that tend to systematically overestimate risk of pesticide transfer are excluded. Knowing the high level of simplification in calculation methods of some indicators, we decided to select both thresholds at a medium level to differentiate the indicators.

2.4.3. ROC test
The Receiving Operating Characteristic (ROC) test consists in comparing indicator outputs with a threshold and calculating, for each value taken by the indicator, the sensitivity and the specificity of the test with regard to this threshold. The sensitivity corresponds to the fraction of true positive situations (the indicator shows a high transfer risk of a given a.i. and the transfer of the a.i. exceeded the threshold) and specificity corresponds to the fraction of true negative situations (the indicators shows a low transfer risk of a given a.i. and the transfer of the a.i. does not exceed the threshold), (Makowski et al., 2009). The next step consists in plotting sensitivity vs. (1 - specificity) values for each indicator value to obtain the ROC curve and to calculate the area under the curve (AUC). The AUC is an assessment of the predictive quality of an indicator: for a perfect indicator, the value of the AUC should be 1, whereas an indicator with an AUC value below 0.5 does not perform a random draw (Supplementary Materials: Table A7). The thresholds selected for this test were: 0.1 µg/L (threshold of drinkable water) and 1 µg/L (half of the threshold for potable water (2 µg/L)) for the weighted average concentration (cmax) and 100 mg/ha for the total flow (ftot).

3. Results

3.1. Preliminary analysis

3.1.1. Analysis of data variability
The results of the indicators calculated for the 4 sites show a satisfactory variability (Supplementary Materials: Table A4). Their variation coefficients vary between 11% and 320% and are higher than 20%, except for the I-PHY1 and DEXiPM-LT indicators on the La Jaillière site (runoff) and DAEGv on the La Jaillière site (drainage) (Supplementary Materials: Table A4). For instance, the I-PHY2v indicator expressed on a scale between 0 and 10, ranges between 2.73 and 10 with a median value of 6.80 on the Jaillière site (drainage). The variation coefficients of the measured variables are much higher than those of the indicators, ranging between 56% and 743%. This is mainly due to their asymmetric distributions, with a majority of points equal to 0. This asymmetry is also demonstrated by the wide gap between the median and the average, the former being much lower than the latter (for example, respectively 0.01 and 0.67 µg/L on the La Jaillière site (drainage)).
3.1.2. Preliminary analyses: Correlation between indicators’ outputs

Several correlations are observed between indicators. Some of them can be explained by a similarity in the construction method (see Fig. 2): for instance the correlation coefficient value ($r$) between I-PHY1 and DAEG is 0.76 at the La Jaillière site (drainage) and $r = 0.95$ at the Geispitzen site. I-PHY1 and ARTHUR have values $r = 0.66$ at the Jaillière site (drainage) and $r = 0.76$ on the Magneraud site. A good relationship is also observed between the indicators having a similar degree of complexity (see Fig. 2). On the one hand, for the simplest ones we can note QSA and EQ ($r = 0.78$ at the La Jaillière site (drainage)) and $r = 0.97$ at the Magneraud site and, on the other hand, for the indicators with a higher degree of complexity we can note that Drainage HAIR and EPRIP have a value of correlation coefficient $r = 0.74$ at the Jaillière site and POCEER and Eprip have a value of $r = 0.75$ at the Magneraud site. All results are highly significant (p-value < 0.001).

3.1.3. Preliminary analyses: Correlation between the measured variables

As previously, high significant correlations are observed: i) between the maximum flow and the total flow (0.96 for the La Jaillière (drainage), 0.90 for La Jaillière (runoff), 0.98 for Le Magneraud and 0.98 for Geispitzen) and ii) between the maximum flow and the maximum concentration (0.96 for the La Jaillière (drainage), 0.74 for La Jaillière (runoff), 0.91 for Le Magneraud and 0.21 for Geispitzen) (Supplementary Materials: Table A11). Like for previous tests, $r = 0.44$ and $r = 0.46$. Correlations on average values confirm the good performance of MACRO ($r > 0.98$) and I-PHY1 ($r$ between 0.69 and 0.63), (supplementary material: Table A9).

For the probability test, MACRO yields results for fd1 with correct estimation of 72% and a probability of 73%. Again I-PHY1 shows best performance among indicators with correct estimation between 51 and 64%, and a probability between 62 and 80%. DRAINAGE-HAIR yields correct estimation above 50% only for fd1. The performance for the probability test of the EQ indicator, a simpler indicator than the previous one (Fig. 1) should be noticed with correct estimations of 63% and 51% respectively for fd1 and cmp, and a probability of 77% and 63%.

The ROC test shows that on this site, all indicators have an area under the curve (AUC) above 0.50 and most of them below 0.80 (Supplementary Table: A11). ARTHUR, I-PHY1 and EQ respectively for 5, 4 and 2 out of 7 cases show AUC above 0.80. In any case, this result shows that all indicators perform better than a random draw.

3.2. Comparison of measured data with the indicator results

3.2.1. Indicators assessing vertical transfer by leaching

3.2.1.1. The La Jaillière site (drainage).

Correlations between indicator outputs and the measured data are generally below 0.5 except for the indicators except for fd1 ($r = 0.79$ for cmax, $r = 0.71$, for ftotal and $r = 0.89$ for cmp) (see Table 3). Some indicators based on an operational model or a quantitative model show a correlation coefficient above 0.40: among those based on a qualitative model (fuzzy decision trees) I-PHY1 (respectively $r = 0.43, 0.42, 0.49, 0.41$ for cmax, fmax, total and cmp), ARTHUR for fd1 ($r = 0.42$) and among those based on a quantitative model, DRAINAGE HAIR and EPRIP for ftotal (respectively $r = 0.44$ and $r = 0.46$). Correlations on average values confirm the performance of MACRO (respectively $0.45$ and $0.50$). Like for previous tests, $r = 0.79$).

For the probability test, MACRO shows the best result for fd1, with correct estimations of 73% and probability of 74%; and acceptable results for cmp with correct estimations of 54% and probability of 91%, while DRAINAGE HAIR yields acceptable results for all the variables, with correct estimation between 50% and 55% and probability between 72% and 88%. As for La Jaillère (drainage), EQ also yields acceptable results for all the variables, with correct estimation between 57% and 63% and probability 65% and 83%. Besides, EPRIP for fd1 (50% of correct estimations and 82% of probability) and I-PHY1 for cmp (correct estimation of 54% and probability of 79%) complete the group indicator with at least 50% of correct estimations.

Like for La Jaillière site, the ROC test shows that on this site, all indicators have an area under the curve (AUC) above 0.50 and most of them being below 0.80 except I-PHY2 that show a AUC below the threshold of 0.50 (Supplementary Materials: Table A11). Like for previous tests,

Table 3

<table>
<thead>
<tr>
<th>Indicator</th>
<th>cmax</th>
<th>fmax</th>
<th>ftotal</th>
<th>cmp</th>
</tr>
</thead>
<tbody>
<tr>
<td>HAIR</td>
<td>0.14</td>
<td>92%</td>
<td>7%</td>
<td>0.14</td>
</tr>
<tr>
<td>I-PHY1</td>
<td>0.30</td>
<td>53%</td>
<td>8%</td>
<td>0.16</td>
</tr>
<tr>
<td>SHIRv</td>
<td>0.82</td>
<td>1%</td>
<td>70%</td>
<td>0.21</td>
</tr>
<tr>
<td>I-PHY2</td>
<td>0.22</td>
<td>77%</td>
<td>63%</td>
<td>0.30</td>
</tr>
<tr>
<td>I-PHY3</td>
<td>0.43</td>
<td>84%</td>
<td>5%</td>
<td>0.21</td>
</tr>
<tr>
<td>DRAINAGE-HAIR</td>
<td>0.28</td>
<td>75%</td>
<td>55%</td>
<td>0.35</td>
</tr>
<tr>
<td>ARTHUR</td>
<td>0.42</td>
<td>82%</td>
<td>45%</td>
<td>0.29</td>
</tr>
<tr>
<td>I-PHY4</td>
<td>0.36</td>
<td>80%</td>
<td>64%</td>
<td>0.44</td>
</tr>
<tr>
<td>I-PHY5</td>
<td>0.24</td>
<td>84%</td>
<td>4%</td>
<td>0.15</td>
</tr>
<tr>
<td>POCEER</td>
<td>0.04</td>
<td>79%</td>
<td>48%</td>
<td>0.08</td>
</tr>
<tr>
<td>EPRIP</td>
<td>0.35</td>
<td>61%</td>
<td>43%</td>
<td>0.39</td>
</tr>
<tr>
<td>DRAINAGE-HAIR</td>
<td>0.31</td>
<td>77%</td>
<td>54%</td>
<td>0.31</td>
</tr>
<tr>
<td>MACRO</td>
<td>0.16</td>
<td>73%</td>
<td>17%</td>
<td>0.79</td>
</tr>
</tbody>
</table>

Legend:

- $r$: Correlation coefficient
- Correct estimation
- $+0$:
  - $[0.4, 0.45]$
  - $[0.45, 0.5]$
  - $[0.5, 0.55]$
  - $[0.56, 0.6]$
  - $[0.6, 0.7]$
  - $[0.7, 0.8]$
  - $[0.8, 0.9]$
  - $[0.9, 1.0]$
Exceedance of the threshold of 0.1 μg/L; cmax: maximum concentration; fmax: maximum flow; ftotal: cumulated flow and cmp: average weighted concentration (All the results not in italic have a p value < 0.01).

### Table 4
Predictive quality of the indicators on the Le Magneraud site: correlation coefficient, probability (correct and overestimation) and correct estimation (see Fig. 2) with fd1: frequency of exceedance of the threshold of 0.1 μg/L; cmax: maximum concentration; fmax: maximum flow; ftotal: cumulated flow and cmp: average weighted concentration (All the results not in italic have a p value < 0.01).

<table>
<thead>
<tr>
<th>r</th>
<th>Correct estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;0.4</td>
<td>[40, 50]</td>
</tr>
<tr>
<td>[0.4; 0.45]</td>
<td>[50, 60]</td>
</tr>
<tr>
<td>[0.50; 0.55]</td>
<td>[60, 70]</td>
</tr>
<tr>
<td>&lt;0.55</td>
<td>&gt;70</td>
</tr>
</tbody>
</table>

DRAINAGE HAIR has an AUC above 0.80 in 6 out of 7 cases while ARTHUR and SIRIS for one case.

#### 3.2.2. Indicators assessing horizontal transfer by runoff

3.2.2.1. The La Jaillière site (runoff). Overall correlations are lower than for drainage except for DAEG (r = 0.57 for fd1 (only notable result for this indicator)). SYNOPs for cmax, fmax, and ftotal respectively r = 0.47, 0.43, 0.49, I-PHY1 and I-PHY2 for fd1 (respectively r = 0.44 and r = 0.49) show correlation coefficient above 0.40 (Table 5). If average values are considered (supplementary material: Table A10), EPRIP2 shows correlation between 0.62 and 0.64 for 3 variables out of 4 (cmax, fmax and ftotal, and cmp: average weighted concentration (All the results not in italic have a p value < 0.01).

The high values of probability highlight the overestimation of the transfer by many indicators, for example SYNOP showing probability between 82 and 90% while correct estimation between 20% and 31%. The ROC test shows results slightly lower than on the previous sites. On this site, all indicators but ARTHUR in one case have an AUC above 0.50, and all are below 0.75 (Supplementary Materials: Table A12).

3.2.2.2. The Geispitzen site (runoff). Like for the La Jaillière site (runoff), correlations are lower on this site than on the jaillière (drainage) and Le Magneraud sites. The exceptions are EPRIP2 having a correlation coefficient r = 0.51 for fd1, r = 0.41 for cmax and r = 0.57 for cmp, and POCER having a correlation coefficient r = 0.41 for fd1 and r = 0.42 for cmp (Table 6). For the probability test, none of the indicators yields >50% of correct estimations. The ROC test shows results lower than on the previous sites. On this site, five indicators yield results with an AUC below 0.50: TFI for two cases with respect cmp, SIRIS for cmax, and

### Table 5
Predictive quality of the indicators on the La Jaillière (runoff) site: correlation coefficient, probability (correct and overestimation) and correct estimation (see Fig. 2) with fd1: frequency of exceedance of the threshold of 0.1 μg/L; cmax: maximum concentration; fmax: maximum flow; ftotal: cumulated flow and cmp: average weighted concentration (All the results not in italic have a p value < 0.01).

<table>
<thead>
<tr>
<th>r</th>
<th>Correct estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;0.4</td>
<td>[40, 50]</td>
</tr>
<tr>
<td>[0.4; 0.45]</td>
<td>[50, 60]</td>
</tr>
<tr>
<td>[0.50; 0.55]</td>
<td>[60, 70]</td>
</tr>
<tr>
<td>&lt;0.55</td>
<td>&gt;70</td>
</tr>
</tbody>
</table>
Table 6

Predictive quality of the indicators on the Gesipretos site: correlation coefficient, probability (correct and overestimation) and correct estimation (see Fig. 2) with fd1: frequency of exceedance of the threshold of 0.1 μg/L; Cmax: maximum concentration; fmax: maximum flow; ftotal: cumulative flow and cmp: average weighted concentration. (All the results not in italic have a p value > 0.01).

<table>
<thead>
<tr>
<th>fdx</th>
<th>cmax</th>
<th>fmax</th>
<th>ftotal</th>
<th>cmp</th>
</tr>
</thead>
<tbody>
<tr>
<td>QAI</td>
<td>0.29</td>
<td>72%</td>
<td>20%</td>
<td>0.33</td>
</tr>
<tr>
<td>TFI a</td>
<td>0.14</td>
<td>70%</td>
<td>15%</td>
<td>-0.06</td>
</tr>
<tr>
<td>NRI s</td>
<td>0.3</td>
<td>92%</td>
<td>33%</td>
<td>0.23</td>
</tr>
<tr>
<td>SIRS s</td>
<td>0.07</td>
<td>50%</td>
<td>23%</td>
<td>-0.01</td>
</tr>
<tr>
<td>EQ s</td>
<td>0.3</td>
<td>50%</td>
<td>33%</td>
<td>0.23</td>
</tr>
<tr>
<td>ADSCOR CT</td>
<td>0.24</td>
<td>35%</td>
<td>15%</td>
<td>0.17</td>
</tr>
<tr>
<td>ADSCOR LT</td>
<td>0.4</td>
<td>27%</td>
<td>13%</td>
<td>0.18</td>
</tr>
<tr>
<td>DEXPIMs low tox</td>
<td>0.27</td>
<td>67%</td>
<td>30%</td>
<td>0.19</td>
</tr>
<tr>
<td>DEXPIMs high tox</td>
<td>0.19</td>
<td>95%</td>
<td>33%</td>
<td>0.31</td>
</tr>
<tr>
<td>DAE</td>
<td>0.3</td>
<td>67%</td>
<td>30%</td>
<td>0.22</td>
</tr>
<tr>
<td>ARTHUR s</td>
<td>0.27</td>
<td>67%</td>
<td>33%</td>
<td>0.38</td>
</tr>
<tr>
<td>1-PHY1 s</td>
<td>0.32</td>
<td>37%</td>
<td>28%</td>
<td>0.21</td>
</tr>
<tr>
<td>1-PHY2 s</td>
<td>0.31</td>
<td>52%</td>
<td>50%</td>
<td>0.34</td>
</tr>
<tr>
<td>POCER s</td>
<td>0.41</td>
<td>54%</td>
<td>21%</td>
<td>0.24</td>
</tr>
<tr>
<td>EPRIP s</td>
<td>0.3</td>
<td>54%</td>
<td>30%</td>
<td>0.23</td>
</tr>
<tr>
<td>EPRIP 2 s</td>
<td>0.51</td>
<td>70%</td>
<td>30%</td>
<td>0.41</td>
</tr>
<tr>
<td>SYNOBS</td>
<td>0.36</td>
<td>97%</td>
<td>29%</td>
<td>0.17</td>
</tr>
</tbody>
</table>

4. Discussion

Overall, the correlation tests show correlations below 0.58 between the different measured variables and the indicator outputs, except the MACRO model on the La Jaillière (drainage) site. The test on the average values of the indicators outputs and the measured variables allows us to pass over heterogeneity of soil conditions between plots and climatic variability between years of the experimentation period. In this case, the MACRO model also yields satisfying results on the Le Magneraud site, and other indicators such as DRAINAGE - HAIR, I-PHY1, EPFRP reach correlations until 0.83 in some cases. Complementary tests such as the probability test recommended by Bockstaller and Girardin (2003) and adapted according to the method of Dubus and Brown (2002), confirm the trends while the ROC tests show the ability of a majority of the indicators to assess the risk to one threshold.

More precisely, the correlation test and the probability test show that the best (the least bad) results are obtained with indicators that take into account the plot environment, the topography, the soil, and the practice parameters and integrate them in a simplified model or in an approach based on a meta-model (see Fig. 1). The I-PHY qualitative indicators belong to this category, while the indicators based on two sub-indicators taking into account, on the one hand the environment, and on the other hand the practices, such as DAEG and ARTHUR, give lower results in general, like the simplest indicators based only one variable the dose (QSA IFT-MA). The EQ indicator, although categorized as a simple indicator, is an exception to this rule, and is based on the outputs of the GLEAMS simulation model used in specific conditions (Kovach et al., 1992). These results of indicators based on one variable or on a simple combination of variables were expected and question the relevance of their comparison. It seems that we tested the ability of a knife to cut a steel bar. However, they are used in many multicriteria assessment methods (Eckert et al., 2000; Gomiero and Giampietro, 2001; Häni et al., 2003) and by regulatory authorities to monitor the evolution of pesticide management (e.g. TFI for the Ecophyto plan). When no other elaborate indicator is available, these simple indicators are even used to assess environmental risk. Our results should serve to discourage this misuse of simple indicators assessing management. On the other side of the gradient of complexity, the mechanistic model MACRO (Larso et al., 2005) integrating precisely the processes of vertical transfer (Casara et al., 2012) logically yields the best results, although it is not always the most complex prediction tool that are the most accurate (Makowski et al., 2009). Although MACRO was not calibrated, these results were obtained for one set of parameters over seven for La Jaillière and four for Le Magneraud site, with soil description of Avralis - Institut du Végétal, and the automatic chain of calculation developed by Agrosolution. (See Supplementary Material: Table A3). Thus, it cannot be considered as a general validation of MACRO. Two indicators based on this model (I-PHY2 and DRAINAGE HAIR) obtain different results. These differences can be explained by their design: DRAINAGE-HAIR was derived from a classical metamodelling approach using a statistical relationship (Strassemeyer et al., 2003) while I-PHY2 was designed with neurofuzzy supervised learning from MACRO simulations for frequency climatic data, more exactly the results for 8 years out of 10 (Lindahl and Bockstaller, 2012).

These poor results for the predictive capacity of most indicators, even the most elaborate is explained by several simplifications. First, actual daily climatic data and water status of soil are not included in the calculation method for the sake of feasibility because these data are difficult to obtain. It can be also partly explained by the construction design of most of the indicators which have been elaborated to predict a risk of pesticide transfer before their application while mechanistic model as MACRO are designed to calculate real transfers with these climatic data. This also allows to pass over the effect of random variation of climate that may hide all the variations of the other variables, for example, the efforts made to change the practices. The period or the date of application, thereby the water status of the soil when pesticides are applied, is another relevant variable, according to a number of works (Real et al., 2005), that is rarely taken directly into account by a majority of indicators (Supplementary material: Table A1), except by I-PHY2. In addition, another explanation lies in the simplifications that were made in our study and that maximize the risk defined by these indicators (see...
Lastly, for the user, this study provides results on the predictive quality of indicators and the MACRO model. Indicators focusing only on properties of active ingredient showed a very low predictive quality, so that their use cannot be recommended. In addition, they invite to an unique advice of a.i. substitution that may lead to pest resistance if the same family of pesticide is advised (Moss et al., 2007). The use of the MACRO model can be recommended with an adapted set of parameters like in this study. When it is not possible to implement it, some more elaborate indicators like I-PHY, EPRIP show the less poor results can be an alternative with much lower cost. All these recommendations are based on the results we obtained from a dataset which is quite unique by its size but remains still limited. In any case, the famous quote “All Models are false, but some are useful” (Box, 1976) could also be applied to indicators.

Acknowledgment

This study started from discussions from the working group “multicriteria evaluation” of the GIS Grandes Cultures À Hautes Performances Economiques et Environnementales (GC-HP2E). The EQUIPE project was supported by the research program “for and on Ecophyto” (Ecophyto PSPE) of the French Ministry in charge of Agriculture by intermediate of ONEMA for 65%. It benefited from complementary financial support from the GIS GC-HP2E and the Water Agency Rhin-Meuse for 35% of the total. Authors are thankful to Chloé Schneller for the preparation of the comparison work and to Alain Dutertre (Arvalis-Institut du Végétal) for the dataset of experimental data. The authors are thankful to Rémé Vuillemin, assistant professor in the University of Strasbourg for his attentive reviewing of English.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.scitotenv.2017.06.112.

References
