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.

Science of the Total Environment 605–606 (2017) 655–665

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

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

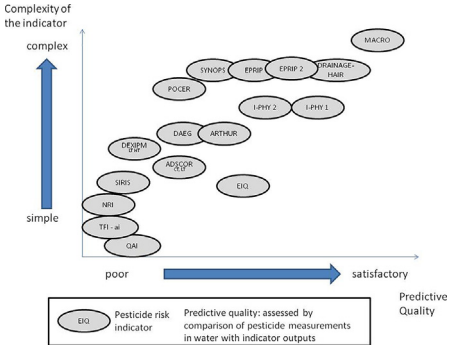
Frédéric Pierlot [a,b,*], Jonathan Marks-Perreau [c], Benoît Réal [c], Nadia Carluer [d], Thibaut Constant [e], Abdeljalil Lioeddine [e], Paul van Dijk [f], Jean Villerd [a], Olivier Keichinger [g], Richard Cherrier [b], Christian Bockstaller [h]

[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] IRSTEA, UR Aquatic 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

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 8 March 2017
Received in revised form 13 June 2017
Accepted 13 June 2017
Available online xxxx

Editor: D. Barcelo

Keywords:
Pesticide transfer
Groundwater
Surface water
Predictive quality assessment

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

* Corresponding author at: Chambre Régionale d'Agriculture Grand Est, Pôle Recherche Développement et Innovations, France.
E-mail addresses: frederic.pierlot@univ-lorraine.fr, frederic.pierlot@grandest.chambagri.fr (F. Pierlot).

http://dx.doi.org/10.1016/j.scitotenv.2017.06.112

environmental aspects yielded comparatively better results. The numerical simulation model MACRO showed much better results than those for indicators. These results will be used to help stakeholders to appropriately select their indicators, and will provide them with advice for possible use and limits in the interpretation of indicator outputs.

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 those 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)

$$fd1 = n1_{ijk}/n_{ijk} \quad \text{(Eq.1)}$$

with $n1_{ijk}$: 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)

$$cmax = MAX(c_{ijk}) \quad \text{(Eq.2)}$$

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

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

$$fmax = MAX(f_{ijk}) \quad \text{(Eq.3)}$$

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.

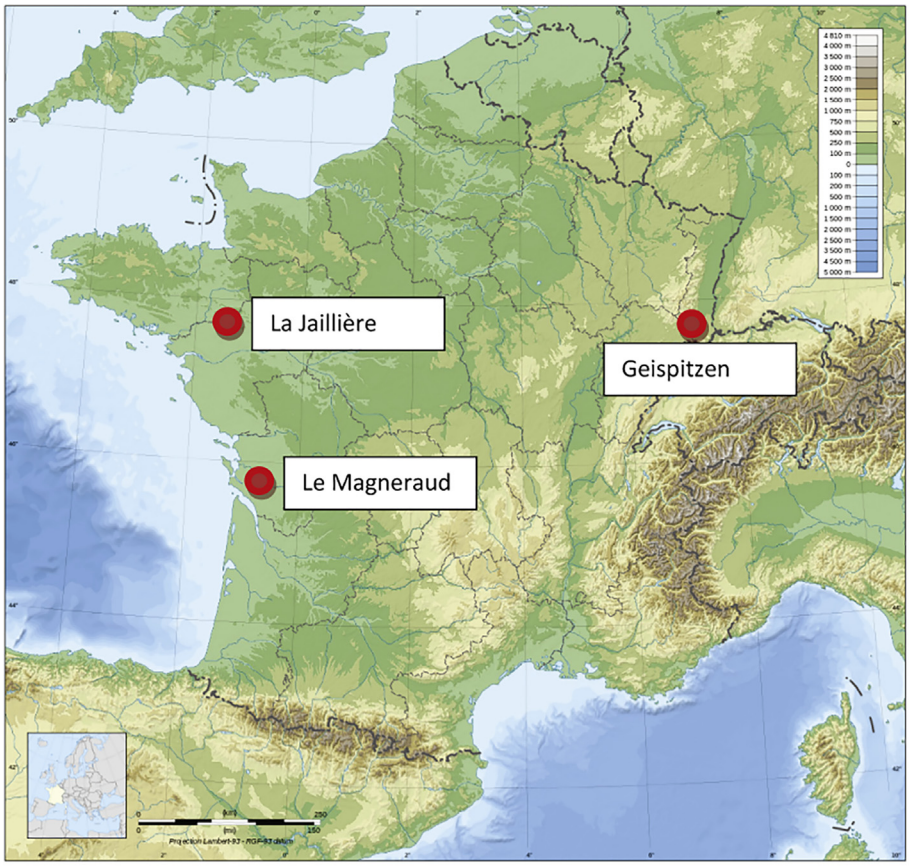


Fig. 1. Physical location of sampling stations.

iv) cumulated flux of active ingredient in mg/ha (ftotal) during the measurement period

$$ftotal = \Sigma(f_{ijk}) \quad \text{(Eq.4)}$$

v) weighted average concentration on the period in μg/L (cmp)

$$cmp = \Sigma c_{ijk}.w_{jk}/\Sigma w_{jk} \quad \text{(Eq.5)}$$

The measurements were performed at field level for the sites of La Jaillière and Geispitzen and for lysimeter of 1 m^2 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 m^2 surface, with no vertical walls and no soil shuffle. The boxes are integrated in 60 m^2 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).

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

Active ingredient	Dosage (g/ha)	Koc (mL/g)	Field DT50 (days)	Application period	Crop types	Site of study
Aclonifen (H)	300–2400	7126	80.4	Spring/fall/winter	M/SP/WP/F	La Jaillière/Le Magneraud
Alachlor (H)	2160–2328	124	14	Spring	M	Geispitzen
Atrazine (H)	250–750	100	29	Spring	M	Geispitzen
Bentazon (H)	261–1740	51.5	10	Spring/winter	M/SP/WP	La Jaillière/Le Magneraud
Bromoxynil (H)	62–400	173.5	5.56	Spring/fall/winter	WW/M	La Jaillière/Le Magneraud/Geispitzen
Chlorothalonil (F)	375–2250	850	44	Spring/winter	WW/SP/WP	La Jaillière/Le Magneraud
Diflufénicanil (H)	20–187	3416	415	Fall/winter	WW	La Jaillière/Le Magneraud
Dmta-p (H)	720–1008	227	7	Spring	M	La Jaillière/Le Magneraud
Epoxiconazol (F)	25–87	1073	116.8	Spring/winter	WW	La Jaillière/Le Magneraud
Glyphosate (H)	480–1080	21,699.44	31.5	Spring/fall/winter	WW/M/WP/CIPAN	La Jaillière
Isoproturon (H)	500–1500	122	22.5	Fall/winter	WW	La Jaillière/Le Magneraud
Mesotrione (H)	30–150	109	5	Spring	M	La Jaillière/Le Magneraud/Geispitzen
S-metolachlor (H)	983–1646	200	21	Spring	M	La Jaillière/Le Magneraud/Geispitzen
Metsulfuron-méthyl (H)	5–30	39.5	31.97	Spring/winter	WW	La Jaillière/Le Magneraud
Nicosulfuron (H)	20–30	20.7	19.3	Spring	M	La Jaillière/Le Magneraud/Geispitzen
Pendimethalin (H)	250–800	15,744	99.17	Spring/fall/winter	WW/SP/WP	La Jaillière/Le Magneraud
Prochloraze (F)	315–450	2225	345.5	Spring/winter	WW	La Jaillière/Le Magneraud
Prosulfocarb (H)	800–3200	1693	9.8	Fall/winter	WW	La Jaillière
Prosulfuron (H)	3.6–15	16.67	16.44	Spring	M	La Jaillière/Le Magneraud/Geispitzen
Tau-fluvalinate (I)	48–72	504,123	90.8	Spring/fall	WW/SP	La Jaillière

(H): Herbicide/ (I): Insecticide/ (F): Fungicide.
M: maize/ SP: spring peas/ WP: winter peas/ F: fababeans/ WW: winter wheat.

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 hills are covered with loess-derived soils of silt loam texture overlying Oligocene molasses and marls. A sloping field (5%) of about 9 ha was divided into 3 bordered fields with measuring flumes and automatic water samplers and the down slope borders just upslope of a ditch draining catchment runoff. The crop system rotation was only composed by maize except for one year, when soybeans were sown. Water samples were taken as a function of runoff volume. Data were collected over the period of 2001 to 2012 but only during the growing season of corn maize. The site was equipped with a weather station owned by Arvalis - Institut du Végétal. Surface runoff occurrence is very irregular in this area and often of rather short duration. As a result, intervals between recorded events on overland flow and pesticides applications were very variable, going from a few days to 3 years. The used database consists of 40 treatments followed by a runoff event between 2000 and 2009, corresponding to 8 used active ingredients (Table 1). When no runoff were measured, it was not possible to distinguish between cases with no runoff and cases with a sampling problem, so that these data were not kept in the dataset.

2.2. Selection of indicators

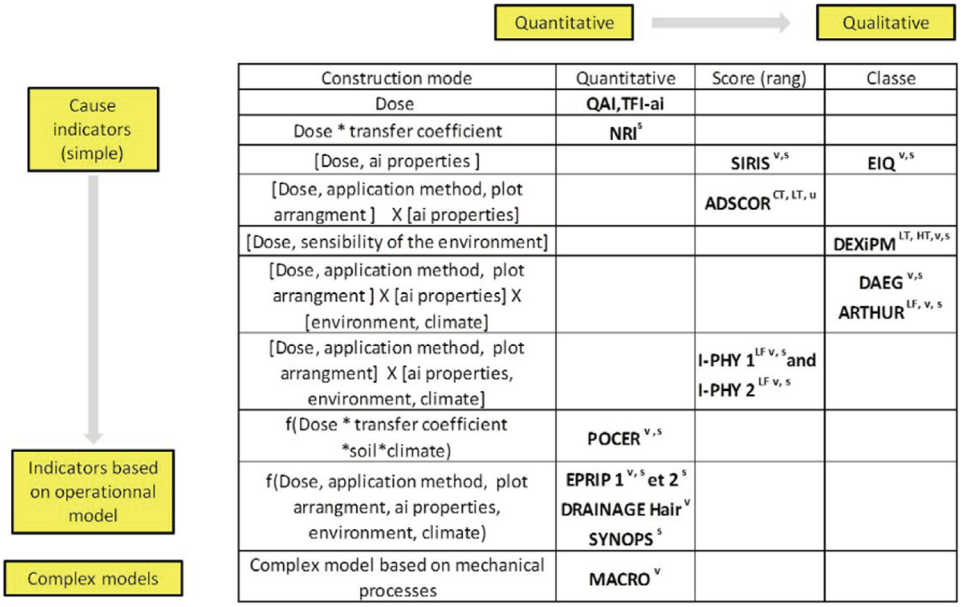
The indicators tested in this study cover the whole gradient of complexity of existing indicators (Bockstaller et al., 2015) which are available to researchers, farmer's advisers and water managers. The set of selected indicators include: i) indicators based on management data (TFI, QAI) using only data 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 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 SYNOPS (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 not need anyone for their calculation. Thus, we also integrated the model MACRO (Larsbo 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 (Microsoft ®) 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



Construction mode	Quantitative	Score (rang)	Classe
Dose	QAI,TFI-ai		
Dose * transfer coefficient	NRI[s]		
[Dose, ai properties]		SIRIS[v,s]	EIQ[v,s]
[Dose, application method, plot arrangment] X [ai properties]		ADSCOR[CT, LT, u]	
[Dose, sensibility of the environment]			DEXiPM[LT, HT, v, s]
[Dose, application method, plot arrangment] X [ai properties] X [environment, climate]			DAEG[v,s] ARTHUR[LF, v, s]
[Dose, application method, plot arrangment] X [ai properties, environment, climate]		I-PHY 1[LF v, s] and I-PHY 2[LF v, s]	
f(Dose * transfer coefficient *soil*climate)	POCER[v,s]		
f(Dose, application method, plot arrangment, ai properties, environment, climate)	EPRIP 1[v,s] et 2[s] DRAINAGE Hair[v] SYNOPS[s]		
Complex model based on mechanical processes	MACRO[v]		

Fig. 2. Synthetic presentation of the method of calculation for the indicators. QSA: quantity active substance; TFI-ai: calculation of the TFI for each application of the active ingredient with the same dose reference values as for the TFIai in course of development; DEXiPM: sub-transfer indicators linked to the use of plant protection products of the DEXiPM model (Pelzer et al., 2012). I-phy2: new version of the indicator I-PHY (I-phy1), Lindahl and Bockstaller (2012) for vertical transfer (v) Wohlfart (2008) for transfer of surface (s)). DRAINAGE-hair: metamodel 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, [v] sub-indicator for vertical transfer (the Jaillière drainage and Le Magneraud infiltration), [s]: sub-indicator for surface transfer (the Jaillière site - runoff by saturation and Geispitzen - hortonien runoff), [LT] and [HT]: low and high toxicity for DEXiPM.

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 directly with the authors for questions of details. For the ARTHUR indicator, the author did not give us access to the detailed calculations, so that we were forced to implement the original software of the tool.

The MACRO model (5.2 version) was parameterized by means of respectively 7 and 4 sets of parameters for the La Jaillière and Le Magneraud sites. They were based either on pedotransfer functions that were available in the model or on the pedotransfer Footprint functions (Centofanti et al., 2008), (See Supplementary Materials: 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 implementation situation with data available in a national database, without any calibration. The outputs of the model used for comparison are fd1, cmax, ftotal and cmp (see Section 2.1 for more details).

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, because 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, K_{oc} 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 provides us with 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 database, 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 K_{oc} and DT50 settings on the correlations between measured variables and indicator 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 coefficient of 0.22), taking results obtained with SIRIS values as a reference. This was mainly due to the high weight on the results of the GUS variable aggregating K_{oc} 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 variability 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%. (Supplementary Materials: Table A4). Then we performed several tests comparing measured variables with indicator outputs, described previously with indicators or MACRO outputs. These tests of decreasing power following recommendations made by Bockstaller and Girardin (2003) and Bockstaller et al. (2008), and the example set by Brown et al. (2002)

were: (i) correlation tests to identify linear relationship, (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 Magneraud 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 (r^2) 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 relation between transfer measurements and indicator outputs. Measured variables were divided into five classes that were equivalent for fd1: 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 pluviometry. 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 (cmp), 1 and 2 μg/l) for the maximal concentration (cmax) and 100 mg/ha for the total flow (ftotal).

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)).

Table 2
Example of distribution of results of indicators and data measured in 5 classes.

QAI	TFI-ai	EIQ	SIRIS	DEXiPM, DAEG, ARTHUR	ADSCOR CT	ADSCOR LT	I-Phy1	I-Phy 2	EPRIP, POCER, SYNOPS, NRI	% Exceedance of threshold of 0,1 μg/L	Concentration (Cmax et CMP) (μg/L)	Flow (mg/ha)
<10	[0; 0.2]	<2	[0; 20]	[0; 1]	[0; 2]	[0; 4]	[0; 2]	[0; 2]	<0,01	[0; 20]	<0.01	<1
[10; 50]	[0.2; 0.4]	[2; 3]	[20; 40]	[1; 2]	[2; 4]	[4; 8]	[2; 4]	[2; 4]	[0,01; 0,1]	[20; 40]	[0.01; 0.1]	[1; 10]
[50; 250]	[0.4; 0.6]	[3; 4]	[40; 60]	[2; 3]	[4; 6]	[8; 12]	[4; 6]	[4; 6]	[0,1; 1]	[40; 60]	[0.1; 1]	[10; 100]
[250; 1250]	[0.6; 0.8]	[4; 5]	[60; 80]	[3; 4]	[6; 8]	[12; 16]	[6; 8]	[6; 8]	[1; 10]	[60; 80]	[1; 10]	[100; 1000]
≥1250	≥0.8	≥5	[80; 100]	[4; 5]	[8; 11]	[16; 20]	[8; 10]	[8; 10]	≥10	[80; 100]	≥10	≥1000

(See Fig. 2 for abbreviation of the indicators and Section 2.2).

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 et the Geispitzen site. I-PHY 1 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 EIQ (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 the site and POCER 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 Material: Table A8).It is beyond the scope of this article to interpret these correlations.

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 bellow 0.5 except for the model MACRO (Table 3). This yields higher correlation than 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, flmax, ftotal 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 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 fd1. The performance for the probability test of the EIQ 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 material: Table A11). ARTHUR, I-PHY1 and EIQ respectively for 5, 4 and 2 out of 7 cases show ACU above 0.80. In any case, this result shows that all indicators perform better than a random draw.

3.2.1.2. The Magneraud site (percolation). Correlations are lower on the Magneraud site than on the Jaillière site (drainage). Only the indicators DRAINAGE HAIR and DEXiPM have correlation coefficients above 0.40, respectively 0.46 and 0.41; this only for fd1. (Table 4). If average values are considered (supplementary material: Table A9); DRAINAGE HAIR for all variables, MACRO for 2 out of 3 show correlations: r between 0.56 and 0.83 for DRAINAGE HAIR and r = 0.62 and r = 0.81 for MACRO, the results being the best for cmax. Two other indicators yield significant results, SIRIS for cmax (r = 0.56) and I-PHY1 for fmax (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 correction 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 Jaillière (drainage), EIQ 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 ACU below the threshold of 0.50 (Supplementary Materials: Table A11). Like for previous tests,

Table 3
Predictive quality of the indicators on the La Jaillière site (drainage): 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 have a p-value < 0.01).

	fd1			cmax			fmax			ftotal			cmp		
	Corr.	Probability	Correct	Corr.	Probability	Correct	Corr.	Probability	Correct	Corr.	Probability	Correct	Corr.	Probability	Correct
QAI	0,14	92%	7%	0,14	94%	21%	0,15	92%	22%	0,20	89%	22%	0,15	99%	15%
TFI a.i.	0,30	93%	8%	0,16	92%	19%	0,18	92%	21%	0,21	89%	21%	0,19	97%	14%
SIRISv	0,32	86%	12%	0,21	71%	19%	0,22	71%	16%	0,25	67%	14%	0,22	84%	22%
EIQv	0,24	77%	63%	0,28	58%	49%	0,30	59%	51%	0,38	58%	50%	0,29	63%	51%
DEXiPMv	0,37	84%	27%	0,21	66%	29%	0,21	68%	30%	0,25	64%	27%	0,22	78%	33%
DAEGv	0,28	75%	59%	0,35	54%	43%	0,34	55%	44%	0,40	55%	44%	0,33	60%	48%
ARTHURv	0,42	82%	45%	0,29	62%	41%	0,29	65%	44%	0,37	61%	40%	0,28	73%	46%
Iphy1v	0,36	80%	64%	0,43	62%	51%	0,42	64%	53%	0,49	63%	55%	0,41	67%	51%
Iphy2v	0,24	84%	41%	0,15	67%	39%	0,16	68%	40%	0,19	64%	37%	0,13	78%	44%
POCERv	0,04	79%	48%	0,08	58%	36%	0,08	59%	36%	0,09	57%	34%	0,07	68%	44%
EPRIPv	0,36	81%	43%	0,39	60%	34%	0,40	60%	34%	0,458	58%	33%	0,39	71%	42%
DRAINAGE HAIR	0,31	77%	58%	0,31	56%	44%	0,37	68%	42%	0,444	68%	44%	0,28	70%	38%
MACRO	0,16	73%	72%	0,79	67%	40%	NA	NA	NA	0,71	59%	59%	0,89	71%	40%

r
- <0,4
- [0,4; 0,45]
- [0,45; 0,50]
- [0,50; 0,55]
- >0,55

Correct estimation
- <40
- [40; 50]
- [50; 60]
- [60; 70]
- >70

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).

	fd1			cmax			fmax			ftotal			cmp		
	Corr.	Probability	Correct	Corr.	Probability	Correct	Corr.	Probability	Correct	Corr.	Probability	Correct	Corr.	Probability	Correct
QAI	0,32	88%	13%	0,06	87%	17%	0,07	87%	17%	0,07	85%	18%	0,01	94%	16%
TFI a.i.	0,25	93%	14%	0,03	90%	21%	0,02	89%	16%	0,02	87%	17%	0,05	98%	20%
SIRISv	0,40	86%	5%	0,13	83%	21%	0,14	84%	22%	0,15	80%	20%	0,06	92%	15%
EIQv	0,34	83%	63%	0,07	67%	58%	0,09	66%	58%	0,09	65%	57%	0,01	76%	60%
DEXiPMv	0,41	81%	30%	0,18	61%	23%	0,2	64%	27%	0,2	61%	25%	0,1	77%	32%
DAEGv	0,28	90%	5%	0,08	93%	14%	0,1	94%	15%	0,12	92%	17%	0	97%	8%
ARTHURv	0,36	85%	30%	0,09	75%	42%	0,1	76%	40%	0,1	72%	38%	0,04	87%	40%
Iphy1v	0,30	82%	49%	0,03	65%	46%	0,05	65%	47%	0,05	65%	47%	−0,01	79%	54%
Iphy2v	−0,17	80%	39%	−0,14	94%	48%	−0,11	59%	27%	−0,11	57%	25%	−0,09	70%	33%
POCERv	0,03	84%	36%	−0,05	75%	34%	−0,05	76%	37%	−0,05	74%	38%	−0,04	85%	39%
EPRIPv	0,02	82%	50%	−0,06	66%	44%	−0,06	67%	46%	−0,06	65%	44%	−0,03	78%	47%
DRAINAGE HAIR	0,46	85%	50%	0,13	73%	54%	0,14	75%	55%	0,16	72%	53%	0,06	86%	55%
MACRO	0,39	74%	73%	0,19	70%	51%	NA	NA	NA	0,23	60%	9%	0,04	91%	54%

r
<0,4
[0,4; 0,45]
[0,45; 0,50]
[0,50; 0,55]
>0,55

Correct estimation
<40
[40; 50]
[50; 60]
[60; 70]
>70

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, flmax, 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 ftotal) and I-PHY1 for 1 variable fmax: r = 0.64 while lower correlations (r between 0.51 and 0.53) are found for I-PHY2, DAEG and DEXiPM-LT with respect to fmax and for I-PHY2 with respect to cmp. For the probability test, none of the indicators yields >50% of correct estimations. 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 and 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 cmax, SIRIS for cmax,

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 0,1μg/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).

	fd1			cmax			fmax			ftotal			cmp		
	Corr.	Probability	Correct	Corr.	Probability	Correct	Corr.	Probability	Correct	Corr.	Probability	Correct	Corr.	Probability	Correct
QAI	0,09	82%	14%	0,18	85%	24%	0,09	94%	22%	0,09	90%	21%	0,13	94%	20%
TFI a.i.	0,32	87%	20%	0,18	80%	14%	0,2	93%	18%	0,26	89%	19%	0,14	90%	15%
NRI s	0,19	93%	13%	0,28	100%	19%	0,2	100%	8%	0,24	98%	13%	0,17	100%	7%
SIRIS s	0,21	68%	13%	0,26	58%	13%	0,29	74%	19%	0,29	69%	16%	0,26	75%	22%
EIQ s	0,23	72%	40%	0,29	62%	34%	0,23	71%	40%	0,27	70%	40%	0,17	69%	35%
ADSCOR CT	0,25	69%	9%	0,25	62%	17%	0,23	79%	20%	0,25	72%	16%	0,23	79%	26%
ADSCOR LT	0,36	67%	13%	0,21	59%	16%	0,2	73%	22%	0,25	67%	17%	0,17	73%	27%
DEXiPMs low tox	0,31	68%	9%	0,27	58%	12%	0,27	74%	17%	0,25	69%	14%	0,28	73%	19%
DEXiPMs high tox	0,22	81%	10%	0,25	88%	26%	0,24	94%	16%	0,26	90%	16%	0,2	96%	18%
DAEG s	0,57	88%	21%	0,21	85%	21%	0,18	90%	17%	0,2	88%	17%	0,2	93%	17%
ARTHUR s	0,21	92%	16%	0,11	91%	14%	−0,04	92%	6%	−0,08	92%	8%	0,09	95%	8%
I–Phy1 s	0,44	54%	51%	0,32	45%	43%	0,27	50%	47%	0,26	50%	47%	0,38	48%	16%
I–Phy2 s	0,49	64%	34%	0,3	53%	33%	0,27	69%	40%	0,28	64%	37%	0,31	66%	40%
POCER s	0,17	78%	13%	0,2	77%	21%	0,14	90%	21%	0,16	84%	18%	0,13	93%	25%
EPRIP s	0,18	59%	35%	0,14	48%	29%	0,11	58%	35%	0,11	56%	34%	0,1	56%	33%
EPRIP 2 s	0,33	66%	25%	0,33	55%	20%	0,28	67%	24%	0,32	63%	23%	0,2	67%	27%
SYNOPS	0,22	83%	20%	0,47	82%	31%	0,43	90%	23%	0,49	87%	25%	0,31	90%	24%

r
<0,4
[0,4; 0,45]
[0,45; 0,50]
[0,50; 0,55]
>0,55

Correct estimation
<40
[40; 50]
[50; 60]
[60; 70]
>70

Table 6
Predictive quality of the indicators on the Geispitzen site: correlation coefficient, probability (correct and overestimation) and correct estimation (see Fig. 2) with fd1: frequency of exceedance of the threshold of 0,1μg/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).

	fd1			cmax			fmax			ftotal			cmp		
	Corr.	Probability	Correct	Corr.	Probability	Correct	Corr.	Probability	Correct	Corr.	Probability	Correct	Corr.	Probability	Correct
QAI	0,29	72%	20%	0,33	57%	25%	0,24	95%	15%	0,25	87%	18%	0,4	77%	28%
TFI a.i.	0,14	70%	15%	−0,06	60%	20%	0,06	92%	18%	0,07	90%	18%	0,13	77%	23%
NRI s	0,3	92%	33%	0,23	87%	48%	0,25	100%	8%	0,26	97%	15%	0,39	97%	30%
SIRIS s	0,07	50%	23%	−0,01	27%	18%	0,01	85%	23%	−0,02	80%	25%	0,01	62%	30%
EIQ s	0,3	50%	33%	0,23	37%	25%	0,25	67%	38%	0,26	62%	35%	0,39	52%	33%
ADSCOR CT	0,24	35%	15%	0,17	27%	20%	−0,1	75%	33%	−0,12	60%	20%	−0,06	47%	20%
ADSCOR LT	0,1	27%	13%	0,18	12%	8%	−0,28	67%	35%	−0,3	55%	25%	−0,17	37%	18%
DEXiPMs low tox	0,27	67%	30%	0,19	45%	25%	0,2	92%	18%	0,2	82%	15%	0,33	70%	25%
DEXiPMs high tox	0,19	95%	33%	0,31	82%	43%	0,15	97%	8%	0,15	95%	8%	0,26	92%	23%
DAEG s	0,3	37%	20%	0,22	22%	18%	−0,13	77%	35%	−0,15	65%	25%	−0,02	52%	28%
ARTHUR s	0,27	67%	33%	0,38	45%	25%	0,18	92%	18%	0,18	85%	23%	0,12	67%	23%
I–Phy1 s	0,32	37%	28%	0,21	25%	25%	−0,12	75%	38%	−0,15	60%	25%	−0,03	47%	28%
I–Phy2 s	0,31	52%	30%	0,34	45%	35%	0,06	80%	23%	0,04	75%	28%	0,18	60%	28%
POCER s	0,41	58%	21%	0,24	39%	18%	0,28	92%	26%	0,28	76%	13%	0,42	61%	18%
EPRIP s	0,3	54%	30%	0,23	35%	19%	0,19	78%	24%	0,17	65%	16%	0,34	51%	14%
EPRIP 2 s	0,51	70%	30%	0,41	50%	30%	0,32	95%	20%	0,3	85%	20%	0,57	67%	18%
SYNOPS	0,36	97%	29%	0,17	92%	45%	−0,07	97%	5%	−0,09	95%	5%	0,01	95%	21%

r
<0,4
[0,4; 0,45]
[0,45; 0,50]
[0,50; 0,55]
>0,55

Correct estimation
<40
[40; 50]
[50; 60]
[60; 70]
>70

ADSCOR-CT for fmax, ftotal and cmp, ADSCOR-LT for fmax and ftotal, and DAEG for ftotal. However, five indicators yield an AUC above 0.80 for fmax: AIQ, NRI, EIQ, POCER and EPRIP2 (Supplementary Materials: Table A12).

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, EPRIP 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 EIQ 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 (Larsbo 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 Arvalis - 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-PHY 2 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

materials and methods). Nevertheless, when we consider the average results, the indicator's predictive quality increases because climatic variations are smoothed.

The indicators assessing the vertical transfers (V) have globally achieved better results than the indicators assessing the surface transfers (S), even when MACRO is not included in the comparison. This may partly be explained by our data set. The Jaillière site is drained and because of that, runoff is limited (Henine et al., 2012) so that the indicators will tend to overestimate the risk on this site. In addition, the indicators tested do not take into account in an explicit manner the runoff by saturation. The runoff and pesticide transfer by runoff is mainly due on the one hand to topographic and soil and on the other to climatic factors (van der Werf, 1996). For the former, the variability between each plot is quite low (see below), so that the climate is the only determinant factor. At Geispitzen, transfer is due to punctual runoff events in spring and summer, so that the weight of climatic data is even higher. Most of the indicators tackle only topographic and soil data, and do not consider climatic variables except EPRIP, POCER and SYNOPS. POCER assesses runoff by a constant coefficient that cannot address the climatic variability while SYNOPS and EPRIP were based on a statistical function estimating runoff from rainfall data. The function of EPRIP seems to be better adapted to the sites of the study.

As already shown in the previous section, the dataset plays a key role in such study. The effort made in this is to our knowledge unique with >1000 treatments and a broad range of variations in terms of active ingredient and climatic variations between studied years for each site. Even in modeling, an experimental dataset of this size has not been implemented until now, even in the COST 66 action, which is one of the most complete studies from a methodological perspective (Vanclooster et al., 2000). The study of Hardy et al. (2008) with 28 a.i. and 34 lysimeters seems to be comparable to ours, but only covers vertical transfer by percolation. However, our dataset presented a limited range of variations for some variables despite its size. For each transfer pathway, there was actually only one site and within each site the variation of topographic and soil conditions between plots was low. This can be illustrated by two input variables of I-PHY1 synthesizing mainly topographic and soil conditions, respectively the potential of leaching and runoff expressed on a scale between 0 (no potential) and 1 (maximum potential). Their values were respectively: 0 to 0.40 at the Jaillière site (drainage), 0.125 to 0.25 at the Jaillière site (runoff), 0.60, to 0.70 at the Magneraud site, and 0.35 to 0.7 at the Geispitzen site. This low variability may partly explain the poor performance of the indicators, without overseeing the others reasons mentioned before. In any case, further studies should be conducted with data sets of the same quality as those we have used but with a broader range of topographic and soil conditions within each site.

5. Conclusions

From a methodological perspective, this study has shown the interest not only to use a correlation test to assess the predictive quality of indicators assessing the risk of transfers to the waters, but also to implement a series of tests as recommended by Bockstaller and Girardin (2003). Despite the unprecedented effort to constitute a large data study, further comparison have to be achieved with dataset with a broader variability topographic and soil conditions within each site.

The designer who intends to develop a new indicator will be able to use some results of this study, including the absence of consideration of given variables playing a key role in the transfer of pesticides to water as the period application and the water status of soil. The relation between the complexity of the indicator and its predictive quality shows that it is in any case necessary to integrate pesticide properties, application, topographic, soil and climatic. The results obtained with I-PHY show that it is possible to develop a qualitative indicator integrating them with a fuzzy decision tree and showing an acceptable predictive quality.

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 Rémi 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

Bockstaller, C., Feschet, P., Angevin, F., Jan 2015. Issues in evaluating sustainability of farming systems with indicators. OCL 22 (1), D102.
Bockstaller, C., Girardin, P., 2003. How to validate environmental indicators. Agric. Syst. 76 (2), 639–653.
Bockstaller, C., Guichard, L., Makowski, D., Aveline, A., Girardin, P., Plantureux, S., 2008. Agri-environmental indicators to assess cropping and farming systems. A review. Agron. Sustain. Dev. 28 (1), 139–149.
Box GEP, 1976. Science and statistics. J. Am. Stat. Assoc. 71 (356), 791–799.
Brown, C.D., Bellamy, P.H., Dubus, I.G., 2002. Prediction of pesticide concentrations found in rivers in the UK. Pest Manag. Sci. 58 (4), 363–373.
Casara, K.P., Vecchiato, A.B., Lourencetti, C., Pinto, A.A., Dores, E.F.G.C., 2012. Environmental dynamics of pesticides in the drainage area of the São Lourenço River headwaters, Mato Grosso State, Brazil. J. Braz. Chem. Soc. 23 (9), 1719–1731.
Centofanti, T., Hollis, J., Blenkinsop, S., Fowler, H., Truckell, I., Dubus, I., et al., 2008. Development of agro-environmental scenarios to support pesticide risk assessment in Europe. Sci. Total Environ. 407 (1), 574–588.
Chen, H., Pan, M., Pan, R., Zhang, M., Liu, X., Lu, C., 2015. Transfer rates of 19 typical pesticides and the relationship with their physicochemical property. J. Agric. Food Chem. 63 (2), 723–730.
Devillers, J., Farret, R., Girardin, P., Soulas, G., 2005. Indicateurs pour évaluer les risques liés à l'utilisation des pesticides. Tec et Doc, Paris.
Dubus, I.G., Brown, C.D., 2002. Sensitivity and first-step uncertainty analyses for the preferential flow model MACRO. J. Environ. Qual. 31 (1), 227–240.
Eckert, H., Breitschuh, G., Sauerbeck, D.R., 2000. Criteria and standards for sustainable agriculture. J. Plant Nutr. Soil Sci. 163 (4), 337–351.
Feola, G., Rahn, E., Binder, C.R., 2011. Suitability of pesticide risk indicators for less developed countries: a comparison. Agric. Ecosyst. Environ. 142 (3–4), 238–245.
Flury, M., Leuenberger, J., Studer, B., Flühler, H., 1995. Transport of anions and herbicides in a loamy and a sandy field soil. Water Resour. Res. 31 (4), 823–835.
Gomiero, T., Giampietro, M., 2001. Multiple-scale integrated analysis of farming systems: the Thuong Lo Commune (Vietnamese Uplands) case study. Popul. Environ. 22 (3), 315–352.
Gouzy, A., Le Gall, A.C., 2007. Mise à jour des bases de données de l'outil SIRIS - Pesticides et amélioration de la méthode Rapport final de la phase 2 du projet. INERIS (p. 90. Report No.: INERIS-DRC-07-84947-16139A-rapport_Ameliorations_SIRIS-2007-vf).
Grung, M., Lin, Y., Zhang, H., Steen, A.O., Huang, J., Zhang, G., et al., 2015. Pesticide levels and environmental risk in aquatic environments in China — a review. Environ. Int. 81, 87–97.

Gutsche, V., Rossberg, D., 1997. SYNOPS 1.1: a model to assess and to compare the environmental risk potential of active ingredients in plant protection products. Agric. Ecosyst. Environ. 64 (2), 181–188.
Häni, F., Braga, F., Stämpfli, A., Keller, T., Fischer, M., Porsche, H., 2003. RISE, a tool for holistic sustainability assessment at the farm level. International Food and Agribusiness Management Review. 6 (4), 78–90.
Hardy, I., Gottesbüren, B., Huber, A., Jene, B., Reinken, G., Resseler, H., 2008. Comparison of Lysimeter results and leaching model calculations for regulatory risk assessment. Journal für Verbraucherschutz und Lebensmittelsicherheit. 3 (4), 364–375.
Henine, H., Chaumont, C., Tournebize, J., Augeard, B., Kao, C., Nedelec, Y., 2012. LE RÔLE DES RÉSEAUX DE DRAINAGE AGRICOLE DANS LE RALENTISSEMENT DYNAMIQUE DES CRUES: INTERPRÉTATION DES DONNÉES DE L'OBSERVATOIRE «ORGEVAL». Sciences Eaux et territoires 16–23 cahier spécial(2012/III).
Keichinger, O., Benoit, P., Boivin, A., Bourrain, X., Briand, O., Chabert, A., et al., 2013. GUIDE: développement d'un outil d'aide à la sélection d'indicateurs de risques liés à la présence des produits phytopharmaceutiques dans les milieux aquatiques - Mise au point, applications et perspectives. Innovation agronomiques. 28, 1–13.
Kovach, J., Petzoldt, C., Degni, J., Tette, J., 1992. A method to measure the environmental impact of pesticides. New York's Food and Life Sciences Bulletin. 139, 1–8.
Larsbo, M., Roulier, S., Stenemo, F., Kasteel, R., Jarvis, N., 2005. An improved dual-permeability model of water flow and solute transport in the vadose zone. Vadose Zone J. 4 (2), 398.
Levitan, L., 2000. «How to» and «why». Crop. Prot. 19 (8–10), 629–636.
Lindahl, A.M.L., Bockstaller, C., 2012. An indicator of pesticide leaching risk to groundwater. Ecol. Indic. 23, 95–108.
Lopez, B., Ollivier, P., Togola, A., Baran, N., Ghestem, J.-P., 2015. Screening of French groundwater for regulated and emerging contaminants. Sci. Total Environ. 518–519, 562–573.
Makowski, D., Tichit, M., Guichard, L., Van Keulen, H., Beaudoin, N., 2009. Measuring the accuracy of agro-environmental indicators. J. Environ. Manag. 90, S139–S146.
Marks Perreau, J., Real, B., Colart, A.-S., Dutertre, A., Bodilis, A.M., 2013. Transfert de produits phytopharmaceutiques par réseaux de drainage et par ruissellement. 22ème conférence du COLUMA. DIJON, pp. 651–661.
Maud, J., Edwards-Jones, G., Quin, F., 2001. Comparative evaluation of pesticide risk indices for policy development and assessment in the United Kingdom. Agric. Ecosyst. Environ. 86 (1), 59–73.
Moss, S.R., Perryman, S.A.M., Tatnell, L.V., 2007. Managing herbicide-resistant blackgrass (*Alopecurus myosuroides*): theory and practice. Weed Technol. 21 (2), 300–309.
Pelzer, E., Fortino, G., Bockstaller, C., Angevin, F., Lamine, C., Moonen, C., et al., 2012. Assessing innovative cropping systems with DEXiPM, a qualitative multi-criteria assessment tool derived from DEXi. Ecol. Indic. 18, 171–182.
Pervanchon, F., Bockstaller, C., Amiaud, B., Peigné, J., Bernard, P.-Y., Vertès, F., et al., 2005. A novel indicator of environmental risks due to nitrogen management on grasslands. Agric. Ecosyst. Environ. 105 (1–2), 1–16.
Real, B., Dutertre, A., Eschenbrenner, G., Bonnifet, J.-P., Lasserre, D., 2005. Résultats de 10 campagnes d'expérimentation: les transferts de produits phytosanitaires vers les eaux varient selon les types de sol. Perspectives Agricoles. 316, 20–24.
Reus, J., Leendertse, P., Bockstaller, C., Fomsgaard, I., Gutsche, V., Lewis, K., et al., 2002. Comparison and evaluation of eight pesticide environmental risk indicators developed in Europe and recommendations for future use. Agric. Ecosyst. Environ. 90 (2), 177–187.
Richardson, M., 1998. Pesticides - friend or foe? Water Sci. Technol. 37 (8), 19–25.
Stenrod, M., Heggen, H., Bolli, R., Eklo, O., 2008. Testing and comparison of three pesticide risk indicator models under Norwegian conditions—a case study in the Skuterud and Heiabekken catchments. Agric. Ecosyst. Environ. 123 (1–3), 15–29.
Strassemeyer, J., Gutsche, V., Brown, C.D., Liess, M., Schriever, C.A., 2003. WP7/D50 definite version final report aquatic indicators. HAIR Project. OECD (p. 89).
Strassemeyer, J., Gutsche, V., 2010. The approach of the German pesticide risk indicator SYNOPS in frame of the National Action Plan for Sustainable Use of Pesticides. OECD 19.
Trevisan, M., Di Guardo, A., Balderacchi, M., 2009. An environmental indicator to drive sustainable pest management practices. Environ. Model. Softw. 24 (8), 994–1002.
Vanclooster, M., Boesten, J.J.T.I., Trevisan, M., Brown, C.D., Capri, E., Eklo, O.M., et al., 2000. A European test of pesticide-leaching models: methodology and major recommendations. Agric. Water Manag. 44 (1–3), 1–19.
Vilain, L., Boisset, K., Girardin, P., Guillaumin, A., Mouchet, C., Viaux, P., et al., 2008. La méthode IDEA: indicateurs de durabilité des exploitations agricoles: guide d'utilisation. Dijon: Éducagri éd.
van der Werf, H.M.G., 1996. Assessing the impact of pesticides on the environment. Agric. Ecosyst. Environ. 60 (2–3), 81–96.
van der Werf, H.M.G., Zimmer, C., 1998. An indicator of pesticide environmental impact based on a fuzzy expert system. Chemosphere 36 (10), 2225–2249.
Wohlfahrt, J., Colin, F., Assaghir, Z., Bockstaller, C., 2010. Assessing the impact of the spatial arrangement of agricultural practices on pesticide runoff in small catchments: combining hydrological modeling and supervised learning. Ecol. Indic. 10 (4), 826–839.
Wohlfart, J., 2008. Développement d'un indicateur d'exposition des eaux de surface aux pertes de pesticides à l'échelle du bassin versant. [Nancy]: INPL.