Journal of Hydrology 535 (2016) 688–698

Contents lists available at ScienceDirect

Journal of Hydrology

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

Including land use information for the spatial estimation of groundwater quality parameters – 1. Local estimation based on neighbourhood composition

T. Heißerer [a], C.P. Haslauer [b,*], A. Bárdossy [a]

[a] University of Stuttgart, Institute for Modelling Hydraulic and Environmental Systems, Department of Hydrology and Geohydrology, Pfaffenwaldring 61, 70569 Stuttgart, Germany
[b] University of Tübingen, Center for Applied Geoscience, WESS, Hölderlinstr. 12, 72076 Tübingen, Germany

ARTICLE INFO

Article history:
Received 26 May 2015
Received in revised form 31 August 2015
Accepted 25 December 2015
Available online 5 January 2016
This manuscript was handled by Peter K. Kitanidis, Editor-in-Chief, with the assistance of Roseanna M. Neupauer, Associate Editor

Keywords:
Statistics
Land use
Groundwater quality
Categorical variable
Mixed distribution
Inverse problem

SUMMARY

Most groundwater recharge comes from the infiltration of water through the land surface. Data analysis shows that solute concentrations at the water table vary between land use categories and depending on the land use composition within a certain neighbourhood.

Driven by these observations, the goal of this paper is to estimate the solute distribution at a location depending on the composition of land use in the neighbourhood, even though land use information is categorical. This goal is achieved by mixing pure distributions of homogeneous land use according to their frequency of occurrence in the vicinity of, and their distance from an estimation location. These pure distributions are jointly inverted using a maximum likelihood-based approach.

The neighbourhood size is optimized using cross-validation. Measurements below detection limit are included via their probabilities of non-exceedance. A solute-specific, spatially distributed measure of information content of the secondary information is presented. The method is applicable for many types of secondary information and can be used as drift for spatial estimation of the primary variable. This estimation is a local estimation and does not include larger scale spatial information. The information of measurements is included via the optimized concentration distributions for land use groups, not via a model of spatial dependence. The global estimation is described in the companion paper.

1. Introduction

Secondary information can be used to improve the estimate of the primary variable. The objective of this paper is to develop a methodology to estimate a primary variable at unsampled locations using a secondary variable that (1) is not only a point measurement but rather an optimized mixture based on the composition of that secondary variable within a certain neighbourhood and (2) that is of categorical type.

The primary variables of this study are groundwater quality parameters measured at the water table and the secondary variable are land use categories, as land use effects groundwater quality. Despite the obvious necessity, the effects of land use on groundwater quality are analysed relatively sparsely. Foley et al. (2005) pointed out that the effects of land use on groundwater quality exist: they presented a detailed assessment of the possible effects of land use change on food production, freshwater and forest resources, regional climate and air quality, and infectious diseases. Meiyappan and Jain (2012) estimate that in the last 250 years more than half of the global ice-free land has been modified by humans. Hydrogeologically, it is clear that land use must have an impact on groundwater quality and recharge rates, as most groundwater originates from excess rainfall infiltrating through the land surface (Foster and Cherlet, 2014). These authors also point out that groundwater response to land use impacts will usually be gradual and often delayed due to the large storage capacity of most aquifer systems. Lerner and Harris (2009) go as far as assessing the extent and effects of planning tools such as source protection zones on land use. Scanlon et al. (2005) worked on quantifying the impact of land use change on groundwater recharge and quality by detailed analysis of vertical water movement within the unsaturated zone, distinguishing between irrigated and dry land agriculture. Some studies exist that quantify the effects of land use more generally in hydrology, for example on flooding (O'Connell et al., 2007; Bronstert et al., 2002). Groundwater analysis is typically performed using numerical models at the scale of wellhead protection areas (Haslauer et al., 2005).

* Corresponding author. Tel.: +49 (0)7071 29 73081.
E-mail address: claus.haslauer@uni-tuebingen.de (C.P. Haslauer).

http://dx.doi.org/10.1016/j.jhydrol.2015.12.049
0022-1694/

Nomenclature

k	$k = 1,\ldots,K$ observations, censored and not censored	n	$n = 1,\ldots,N$ location in the vicinity of target
i	$i = 1,\ldots,I$ observations, not censored	l_k	land use at location x_k
j	$j = 1,\ldots,J$ observations, censored	d	Euclidean distance
g	$g = 1,\ldots,G$ land use group	α_n	weight at location n
s	$s = 1,\ldots,S$ supporting point	h	kernel density
ϕ_{sg}	weight of group g at supporting point s	H	kernel distribution
γ_{kg}	weight of group g in the vicinity of point k	v_g	frequency of occurrence of group g
x	coordinates of a location k	U	uniformly distributed variable
z	observed value/interpolation quantity	b	constant
0	target/interpolation location		

The methodology demonstrated in this paper utilizes land use as secondary information to enhance the estimation of anthropogenic groundwater quality at unsampled locations. For example, one would expect that nitrate concentrations in groundwater under farm fields are generally larger than under forests due to the excessive application of fertilizers. Groundwater quality parameters are influenced by anthropogenic, biogenic, and geogenic processes. For shallow groundwater systems, anthropogenic processes are dominant, hence the composition of land use is expected to provide useful information for the estimation of groundwater quality parameters close to the land surface. This paper presents methodologies to incorporate information about categorical land use data within the vicinity of an interpolation location to estimate the contaminant distribution at any location where measurements may not necessarily exist. There are two important factors to consider: (1) not only a mean, but a full distribution is estimated at any location, and (2) not only will the land use directly at the interpolation location be used, but the composition of the land use within the vicinity of every interpolation location. This distribution is referred to in this paper as "locally mixed distribution", whereas the distributions of the land use groups are referred to as "global" or "pure" distribution functions. Ultimately, the locally mixed distributions will lead to an improved estimation of spatially distributed parameters relevant for hydrology.

There are two key problems that are solved with the presented methodologies: (1) Land use is categorical information. Based on categories alone it is not possible to derive or estimate the distribution of the primary variable at a location where it was not measured; (2) The composition of the land use in the vicinity of an interpolation location should be considered. That means that the value of the secondary information directly at the interpolation location alone is not sufficient for a good estimate, but instead a neighbourhood with a certain size and associated land use composition should be evaluated. These two basic hypotheses have four main implications: (a) The distribution of the secondary information at an unsampled location must be a mixture of the pure distributions of the categories that exist within a given neighbourhood. The distribution of the secondary information at an unsampled location is then allowed to vary at each location as the composition of the neighbourhood varies; (b) the pure distributions for a given neighbourhood size are unknown and must be estimated (this is an inverse problem); (c) the number of categories of secondary information should be minimized. This number becomes more and more important as larger neighbourhoods are considered: the larger the number of categories, the tougher the inverse problem is to solve; (d) The size and the shape of the neighbourhoods are subject to optimization.

Our hypothesis was that the distribution of a primary variable can be estimated at an unsampled location by a mixed distribution reflecting the composition of the secondary information within the neighbourhood of a certain size and shape. The distributions used for mixing and the ideal neighbourhood size are subject to optimization. For anthropogenic contaminants, these locally mixed distributions represent the contribution of vertical infiltration. This information could be used in a copula-based geostatistical framework together with the horizontal transport component to interpolate groundwater quality.

Intuitively, a certain area with its individual composition of land uses should effect the concentration of a contaminant at a continuous shallow groundwater table at a given point. This composition is different in different parts of the domain, hence the result is a "local" distribution. Because the composition of the land use in the vicinity of a given point is incorporated, it is also called "mixed". The resulting distribution that is influenced by the composition of the secondary information in its vicinity is hence being referred to as "mixed local distribution". For example, the nitrate concentration at an interpolation location that lies within a forest, but is close to a farm field, is expected to be influenced by potentially larger nitrate concentrations under the farm field, and also by the potentially smaller nitrate concentrations under the forest. Such an approach based on mixture distribution has the benefit of smoother and more realistic boundaries. The size and the shape of the neighbourhood are subject to optimization. The underlying distributions used for mixing are "pure" distributions, i.e. distributions of a homogeneous neighbourhood composed of one group of secondary information. These distributions vary with neighbourhood size and are generally unknown and need to be estimated via inversion. Rarely, a sufficient number of measurements exists with large enough uniform neighbourhoods to estimate these distributions.

In this study, the focus is not on censored measurements (e.g. measurements below detection limit), but they are included in the estimation of the marginal distribution. It has been long accepted that censored measurements contain useful information that should be used in (low-dimensional statistical) models – see amongst others Cohen (1976), Helsel and Gilliom (1986), Helsel and Cohn (1988), Liu et al. (1997), Kroll and Stedinger (1999) and Shumway and Azari (2002) for reference.

This paper is structured in the following way: First it is demonstrated based on the data used, that there is considerable variability of contaminant concentrations within land use classes and that differently composed neighbourhoods lead to different contaminant concentrations. Following these motivational statements, the methodology of employing non-parametric distribution functions to describe measurements is presented and extended to incorporate censored measurements (Section 3). Secondly, the process of merging similar distribution functions is shown (Section 4), which is needed to reduce the parameter space. The concept of mixed distributions is explained in Section 5, which is needed when the "pure" distribution functions (pure in the sense of a single land use group within a given neighbourhood size) need to be optimized (Section 6). Finally, tests for determining an ideal

neighbourhood size are shown (Section 7) and the value of the secondary information is evaluated in a spatial sense (Section 8).

2. Data and motivation

This section describes the data that were used for this study and describes the motivation for why the composition of the land use has to be taken into account for estimating a locally mixed distribution function, given the data.

2.1. Data used

The data set used contains groundwater quality parameters which were collected within one recent sampling campaign by the state agency for the environment, measurement, and natural protection of the state of Baden-Württemberg, Germany. For this study, the contaminants nitrate, chloride, and barium were used. Each measurement is representative of the concentration at the groundwater table. There are $n > 2000$ measurements available within the interpolation domain of the state of Baden-Württemberg (36,000 km^2), south-west Germany.

Table 1 lists the groundwater quality parameters with their basic descriptive statistical measures. The arithmetic means differ by a few orders of magnitude. Most parameters exhibit a large skewness, which is caused by a few large measurements. Such extremes are defined in this paper as values whose quantile is larger than 98%. The methods used in this paper are robust enough to not be significantly influenced by these extremes.

Fig. 1 shows the readily available secondary information of 16 land use categories within the state of Baden-Württemberg that is used in this study provided by the state agency for the environment, measurement, and natural protection of the state of Baden-Württemberg, Germany. Each pixel of the map shows the dominant land use category within a 30 m × 30 m area. This land use category was obtained from an analysis of Landsat satellite images and validated (Jacobs, 2001). Coloured scattered dots indicate measurements values of nitrates at their measurement location. Dominant land use within the state is agriculture, whereas most of the observations were taken in extensive grasslands (Fig. 2, Table 2). Besides these two extremes, there is a discrepancy between the frequencies of land use categories occurring within the state and the frequencies of land use categories at which observations were made – a typical bias of observation networks (Fig. 2). The proposed method helps to decrease the negative effects of this bias.

2.2. Motivation

Stationarity assumes that a sample anywhere within a domain is a realization from the same distribution function that is constant over the domain. A model that relies on this assumption is questionable, at least for the given case where we consider anthropogenic contaminants at the water table at the spatial scale of a

Table 1
Descriptive statistical measures of the groundwater quality parameters presented in this paper. Censored measurements are ignored. DL stands for detection limit, max DL for the largest censorship level that occurs for each parameter, *C* for concentration, *n* tot for the total number of measurements, avg for the arithmetic average, stDev for standard deviation, var for variance, skew for skewness.

Descriptive measure	Unit	Barium	Chloride	Nitrate
n tot	[–]	2065	2800	4236
% < DL	[–]	2.90	0.10	5.20
max DL	[mg/L]	0.10	0.50	0.50
Min	[mg/L]	0.01	0.50	0.40
max	[mg/L]	1.00	854.90	179.90
avg	[mg/L]	0.13	34.52	21.17
Median	[mg/L]	0.08	22.30	16.90
stDev	[mg/L]	0.14	50.62	18.06
var	[(mg/L)^2]	0.02	2562.44	326.15
skew	[–]	2.51	6.81	1.55

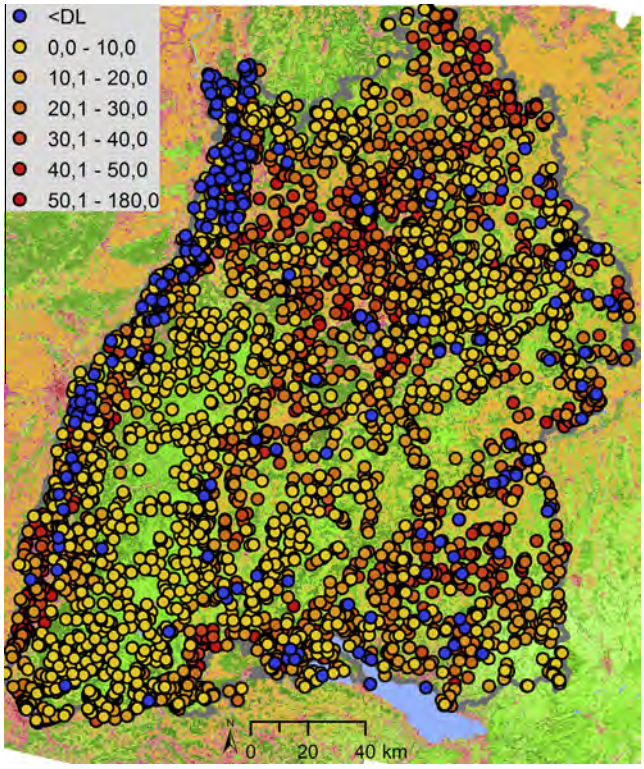

Fig. 1. Map of the domain, the state of Baden-Württemberg in south-west Germany. The shaded colours in the background represent the original 16 land use categories. The coloured dots indicate the measurement locations and the nitrate measurements [mg/L].

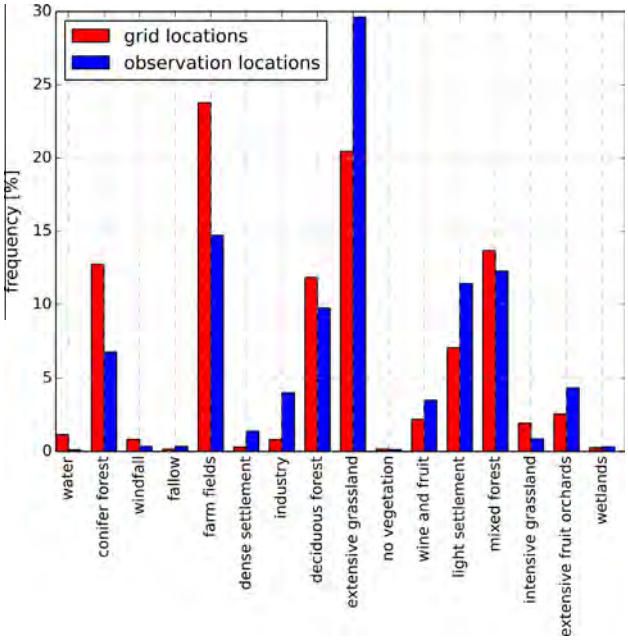

Fig. 2. Relative frequencies of land use categories occurring within the interpolation domain (red bars) and land use categories within which observation locations are placed (blue bars) for nitrate. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Descriptive statistical measures of the observed nitrate concentrations (C) for each land use category. C stands for concentration, n for the number of measurements within this category, stdDev for standard deviation, skew for skewness.

Land use category	Mean [mg/L]	stdDev [mg/L]	skew [–]	n [–]
Water	17.5	22.6	−0.473	5
Conifer forest	11.2	12.3	7.813	287
Windfall	6.6	5.9	2.477	15
Fallow	14.0	11.9	0.712	15
Farm fields	31.8	24.1	2.61	624
Dense settlement	25.7	21.5	2.228	59
Industry	17.4	16.9	4.379	170
Deciduous forest	16.8	13.9	3.312	413
Extensive grassland	20.9	15.0	1.021	1255
No vegetation	21.8	15.3	−0.827	6
Wine and fruit	29.6	20.0	−0.261	148
Light settlement	21.9	19.5	4.152	484
Mixed forest	15.0	13.3	3.619	520
Intensive grassland	20.2	15.9	4.418	37
Extensive fruit orchards	25.7	14.4	−0.485	184
Wetlands	22.7	12.6	−1.318	14

province, where it is known that the distribution of nitrate concentrations under agricultural land is generally ~2–3 times smaller than the nitrate concentration distribution under forested land (Table 2). Even more, not only the mean nitrate concentration varies with land use categories, but there is considerable variability and skewness within each category (Table 2). Generally, the nitrate concentration below forest is low, but large measurements were also observed. This means that the distributions of a contaminant likely will differ, depending on the land use category under which it was measured. Even more, land use categories with smaller mean nitrate concentrations tend to exhibit a larger skewness, indicating that most of the measured concentrations under "wine and fruit" fields are large. This means also that information about the land use category provides useful knowledge about the primary variable groundwater quality, and hence it should be incorporated into a model that estimates solute concentrations in groundwater at locations where no measurement is available.

The spatial variability in spatially correlated groundwater quality data has been explored using the concept of mutual entropy by Mogheir et al. (2004). The mutual entropy (information) between x and y, also called transinformation ($T(x,y)$, Eq. (1)), is interpreted as the reduction in uncertainty in x (solute concentration), due to the knowledge of the random variable y (land use). Transinformation can be calculated empirically on variables on both categorical and rational scale based on marginal entropy ($H(x)$ and $H(y)$, Eqs. (2) and (3)) and joint entropy ($H(x,y)$, Eq. (4)). No distance measures between variables are required. We ensured the comparability of $T(x,y)$ between contaminants by asserting equally probable classes of solute concentrations per land use group.

$$T(x,y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i,y_i)\ln\left[\frac{p(x_i,y_i)}{p(x_i)p(y_i)}\right] \quad (1)$$

$$H(x) = -\sum_{i=1}^{n} p(x_i)\ln(p(x_i)) \quad (2)$$

$$H(y) = -\sum_{i=1}^{n} p(y_i)\ln(p(y_i)) \quad (3)$$

$$H(x,y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i,y_i)\ln(p(x_i,y_i)) \quad (4)$$

Between the contaminants considered in this study, transinformation between land use and the contaminants nitrate ($T = 0.138$) and chloride ($T = 0.258$) is large, whereas for barium ($T = 0.083$) quite small. These transinformation values correspond well with the maps of information content that will be presented in Section 8.

Furthermore, not only does the information of land use effect groundwater quality at a point in space, but the composition of the land use within a certain neighbourhood of a discrete location does. For nitrate, the concentrations under agricultural lands are expected to be different than under forests. At the interface between agricultural land and forest, the concentration should reflect both land uses, according to the proportions of occurrence and with a certain distance weight. Generally, locally mixed distribution reflect the composition of the land use in the neighbourhood.

These effects of the land use composition within the vicinity and the size of the neighbourhood on contaminant concentrations are detectable in the data. The available measurements were subselected based on neighbourhoods with different frequencies of land use groups and with different sizes of neighbourhoods. Fig. 3 shows empirical distributions for nitrate, based on measurements whose neighbourhoods' compositions were differing in their degree of homogeneity. In this study, two kinds of neighbourhoods are shown: (1) almost homogeneous neighbourhoods where 66–100% of the neighbourhood was of one single land use group, and (2) medium heterogeneous neighbourhoods where 33–66% of the neighbourhood was of one single land use group. However, the distributions are not only differing for a different degree of homogeneity of the composition of the neighbourhood, but also for different neighbourhood sizes, as visualized on the three panels of Fig. 3.

The secondary information, land use, does have an effect on the contaminant concentrations: different land uses lead to different concentrations, as reflected by the composition and size of the neighbourhood considered. Hence, the basic assumption of stationarity is not valid under these circumstances. The methodology proposed in this paper does not violate this assumption as it takes two factors into account: the degree of heterogeneity of the composition of the neighbourhood and the size of the neighbourhood.

3. Non-parametric distribution functions

Both the global and the locally mixed distributions are non-parametric distribution functions whose kernel-width is optimized using leave-one-out cross-validation (Hawkins et al., 2003). This method can incorporate censored measurements (i.e. measurements below detection limit), possibly with varying thresholds of censorship. Non-parametric distributions were chosen due to the large number of available measurements which allowed us to stay as close as possible to the data. Both the lower part of the distribution (that includes the censored measurements) and the upper part of the distribution (that contains unusually large values) should be reflected equally well in the estimated distribution function. Censored measurements were included into the likelihood as probabilities of non-exceedance.

The available data sets contain a relatively large number of samples, the measurement values are relatively skewed, and there can be a significant portion of censored measurements (Table 1). All three factors can be taken into account when estimating a distribution function via non-parametric distributions within a maximum likelihood (L) based scheme to optimize the kernel width θ (Eq. (5)) including k not-censored (crisp) and l censored measurements, censored at level z_j^o.

$$L(\theta) = \prod_{i=1}^{k} f(z_i,\theta)\cdot\prod_{j=1}^{l} F\left(z_j^o,\theta\right) \rightarrow \max \quad (5)$$

Due to the large number of available measurements, non-parametric kernel-based distribution functions were not fitted with points of support at each measurement value, but at a significant smaller number of points of support ($n < 100$). By doing so,

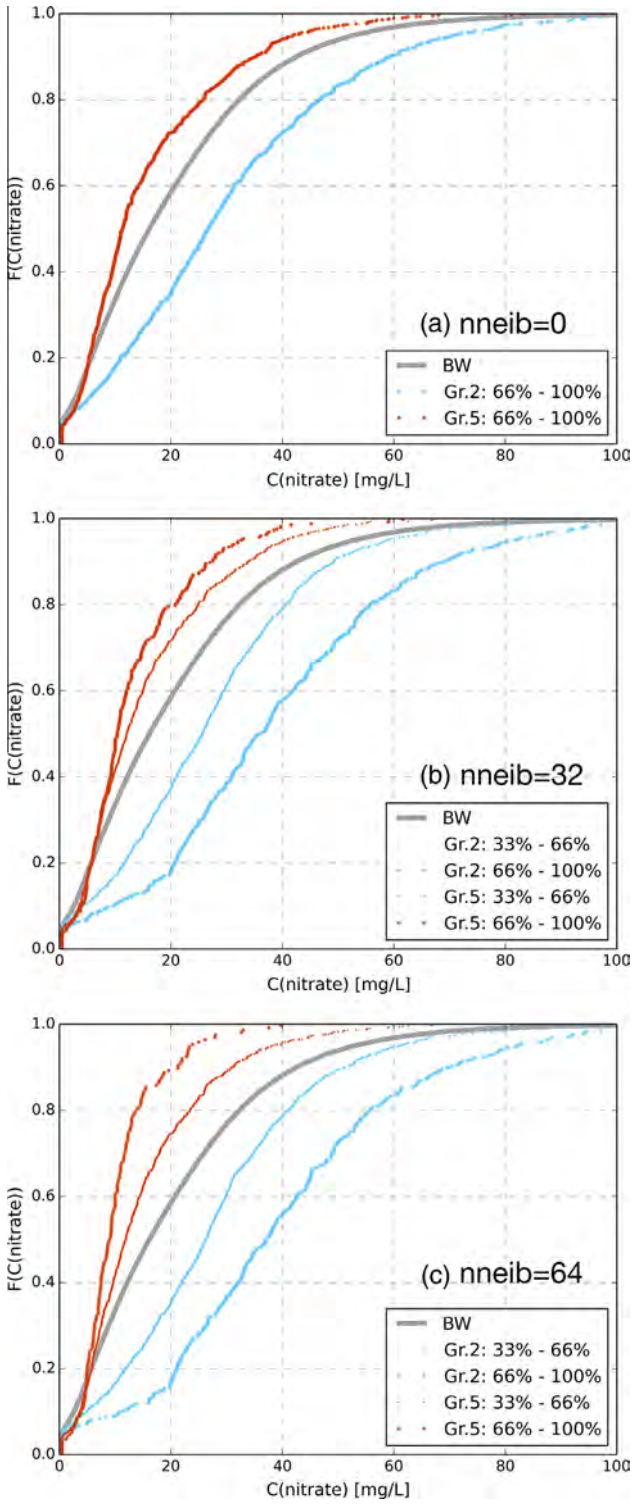

Fig. 3. Empirical distributions of nitrate for different neighbourhood sizes ($nneib \in (0, 32, 64)$) and different degree of land use homogeneity in the neighbourhood (shown are as examples different percentages of groups 2 ("Gr.2", blue, agricultural areas) and 5 ("Gr.5", red, forrest) within a certain neighbourhood). For comparison, the global distribution of Baden-Württemberg (BW) is shown in grey.

the parameter space was minimized. This approach proved advantageous, as only the change in the weights of each point of support needed to be taken into account during optimization at later stages of the proposed method. The chosen approach is stable over a large variety of situations, including cases with a large proportion of censored measurements or skewed distribution of the measurements (small n in Table 1).

This methodology can be applied not only for the entire set of measurements of a given contaminant over the entire state, but also for subsets of measurements of a given contaminant that are measured within a certain land use category.

4. Merging similar distribution functions

With the method described above, a distribution function was estimated for each of the 16 available land use categories. Similar categories were grouped: (a) for reasons of parsimony we did not impose a difference on similar categories, and (b) the sample size in some land use categories was negligible (Table 2). Treating those categories separately would lead to non-meaningful and non-reproducible statements.

Initially, there were 16 distributions based on the land use categories encountered directly at a measurement location. During the process of grouping, the measurements within a pair of land use categories were combined if their distributions were most similar among all other distributions, based on a KS-test with a level of significance of 5%. Table 3 lists the six groups of similar nitrate distributions. The result of the grouping is specific for each contaminant and is assumed to be constant for the interpolation domain.

For each land use group, distributions were fitted to the underlying measurements whose shape varies between different land use groups, and the chosen non-parametric approach leads to irregular shapes (e.g. bimodal behaviour) that would not have been possible with analytic distributions. These land use group-specific distributions are some kind of local distributions, in that they are more specific than the global distribution. However, they are not sufficiently local yet as they do not include information about the composition of the neighbourhood.

5. Locally mixed distribution functions

Knowing that the composition of the neighbourhood effects contaminant concentrations (Section 2), the goal was to develop a model that assigns to an arbitrary location a local distribution function, which takes the composition of the land use in the vicinity into account. The model that reflects such behaviour assumes that the locally mixed distributions are a sum of the pure distributions in the vicinity weighted by the frequency of the land use groups occurring in the vicinity and also the distance relative to the point of estimation. This weighting approach is flexible in that

Table 3
Groups of similar land use categories for nitrate in Baden-Württemberg, based on concentration distributions at measurement locations.

Group ID	Grouped land use categories
0	Dense settlement Light settlement Extensive grassland
1	Industry Others Intensive grassland
2	Farm fields Wine and fruit Extensive fruit orchards
3	Conifer forest
4	Windfall
5	Deciduous forest Mixed forest

it can easily accommodate distance weighting or the inclusion of other types of secondary information. The problem that we do not know the pure distributions remains. But we know that they differ for varying neighbourhood sizes.

The proposed model constructs at each interpolation point x_0 a locally mixed distribution function F_0 which is the weighted sum of the pure group-wise distribution functions F_g of the land use groups that occur within the selected neighbourhood with the associated weights γ_{0g} (Eq. (6)). With this notation, F_g can vary with x. The weights contain a distance weighting factor α_n (Eq. (7)), which in this paper was a weighting proportional to the inverse of the squared distance between each raster-cell i in the neighbourhood and the interpolation location (Eq. (6)). The condition for the weights of the neighbouring points is that they sum to unity (Eq. (8)).

$$F_{x_0}(z) = \sum_{g=1}^{G} \gamma_{0g} \cdot F_g(z) = \sum_{(x_i - x_0) < d_{\max}} \alpha_{x_i} \cdot F_{g(x_i)}(z) \tag{6}$$

$$\alpha_{x_i} \sim b \cdot \exp(d^2) \tag{7}$$

$$\sum_{(x_i - x_0) < d} \alpha_{x_i} = 1 \tag{8}$$

The weights for the mixed local distributions at an interpolation location reflect not only the proportions that are covered by a land use group, but also the distance from the interpolation location. Cases can occur where a certain land use is dominating in a neighbourhood, but a lesser weight is finally assigned to this land use depending on the geometric configuration (Fig. 4). If the composition of the land use in a neighbourhood is considered, the size of the neighbourhood is expressed in units of number of neighbours (*nneib*) of raster cells containing land use information in radial distance. Radii for circular shaped neighbourhoods used commonly in this paper include ~ 1 km ($nneib = 32$), ~ 1.5 km ($nneib = 48$), ~ 2 km ($nneib = 64$), ~ 4 km ($nneib = 128$).

6. Estimation of pure distribution functions via ML

The previous section laid the foundation for the model for the locally mixed distribution functions. This section describes the construction of the underlying pure distributions for neighbourhood sizes larger than zero. These functions are unknown a priori, due to insufficient number of samples in neighbourhoods of different sizes with homogeneous land use composition for different land use groups. Hence, an inverse approach was chosen: via the locally mixed distributions, the underlying pure distributions can be carved out.

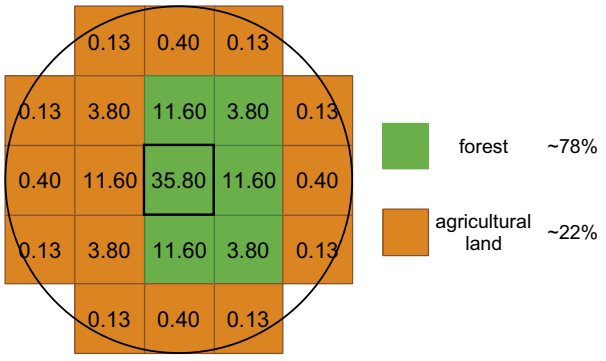

Fig. 4. Sketch to illustrate the concept of locally mixed distributions. Despite the fact that agricultural land ($\sim 71\%$ by area, orange shading) dominates forest in the neighbourhood of the interpolation point (raster cell with thick black outline), the weight factors (shown as percentages in each cell) lead to a dominating weight of the forest ($\sim 78\%$). The sum of all weights in the neighbourhood is 100%. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

In this inversion approach, the F_0 in Eq. (6) must be established. This is a high-dimensional optimization problem due to the fact that there are many measurements and because the pure distributions for all groups of a given contaminant have to be optimized jointly for a given neighbourhood size. As for the estimation of the global distribution function, the approach of using supporting points to construct the non-parametric distribution functions was chosen, in order to maintain a manageable parameter space (as in Section 3).

The first step of the inversion procedure is to establish an initial guess of a pure distribution function for each land use group. A log-normal kernel is optimized for each group of secondary information separately by taking into account point data only. The supporting points of the kernel are chosen uniformly in probability space. Using these initial pure distributions, local distributions are mixed at each location where a measurement was available, using Eq. (6). Based on these distributions, the quantiles of the measurements (F_j) were calculated for censored measurements or the density values for not-censored measurements (f_i), which were needed to calculate the likelihood of the ML-based optimization (Eq. (9)).

With mixed distributions as a basic concept, a mixed likelihood (Eqs. (9)–(12)) is maximized using the Great Deluge Algorithm (Dueck, 1993). The pure distributions were modified until the likelihood was maximal. Within this process, a modification of a pure distribution is achieved by varying the weighting w_{sg} of s supporting points of the non-parametric distribution functions. Initially, the weights w_{sg} were uniform and at any time they are forced to sum to unity (Eq. (12)).

The goal of the optimization routine was to look for a plausible solution. Sets of slightly different weights could potentially lead to similar or even larger likelihoods, but the effect on the distribution functions was considered to be negligible. The mixed likelihood is specific for a specific neighbourhood size. The optimization derives the underlying "pure" distributions for this particular neighbourhood.

Furthermore, this approach allows the inclusion of both "censored" measurements (i.e. measurements below detection limit) and "crisp" measurements (i.e. measurements above detection limit). Crisp measurements are incorporated via the probability density function of the measurement f (Eq. (9)) and censored measurements via their probability of non-exceedance (Eq. (11)). In Eq. (9), j represents censored measurements, i represent crisp measurements, g are the groups of secondary information with an occurrence ratio of γ_{kg} at point k, k are the supporting points weighted with w_{kg}.

$$L(\theta) = \prod_{i=1}^{I} f(z_i, w_{ig}) \cdot \prod_{j=1}^{J} F\left(z_j^o, w_{jg}\right) \rightarrow \max \tag{9}$$

$$f(z_i, w_{ig}) = \sum_{g=1}^{G} \gamma_{ig} \cdot \sum_{s=1}^{S} w_{sg} \cdot h_{hig} \tag{10}$$

$$F\left(z_j^o, w_{ig}\right) = \sum_{g=1}^{G} \gamma_{jg} \cdot \sum_{s=1}^{S} w_{sg} \cdot H_{jgs} \tag{11}$$

$$\sum_{k=1}^{K} w_{kg} = 1 \quad g = 1, \ldots, G \tag{12}$$

6.1. Validation

The plausibility of the model was checked using a similar analysis that served as motivation for including secondary information

from the vicinity of an estimation location (Section 2). Local distributions were mixed using the optimized pure distributions at the locations that fulfilled two criteria: (1) where more than 70% of the composition of the neighbourhood was of a single land use category, and (2) where a measurement was available. These locally mixed distributions were averaged and plotted as solid lines next to the corresponding empirical distributions (Fig. 5). Generally, the locally mixed and empirical distributions fit remarkably well, considering that the pure distributions were not fitted for this scenario. Instead, the fit is general for every location. Groups where not enough measurements with sufficiently pure neighbourhoods were found are not shown. The worst fit occurs for the large values of group two, which may be attributed to an insufficiently not stable optimum in the optimization procedure.

6.2. Examples

A series of figures show optimized distribution functions for nitrate and for three different neighbourhood sizes (left panels on Fig. 6) and maps of the expected value of locally mixed distributions for the corresponding optimized distribution functions (right panels in Fig. 6). These maps do not contain information about the spatial dependence in the sense of a geostatistical model. In the context of this set of papers, the method to produce maps of concentrations by plotting the expected value of locally mixed distributions is referred to as method "A2". This information could be included as secondary information for geostatistical approaches.

Starting at a neighbourhood size of zero and moving to larger neighbourhood sizes, the spatial smoothness of the expected concentrations increases and the range of values increases as the pure distribution functions become more dis-similar. Patterns based on land use become visible, such as larger valleys in the Black Forest (central and southern portion along the western part of the domain). Then, for even larger neighbourhood sizes, the variability in expected values decreases. This shrinkage back to close to original conditions is expected because in the limit of an infinite neighbourhood size, the distributions of the groups should approach the global distribution of Baden-Württemberg.

The effect of land use on barium concentrations (Fig. 7a) is pronouncedly smaller than on nitrate (Fig. 6c, right panel) and chloride concentrations (Fig. 7b). The spatial pattern of locally mixed distributions is similar between nitrate and chloride concentrations, even though large chloride concentrations are locally more confined than in the case of nitrate. The concepts and results of the process of constructing underlying pure distributions and the local mixing is identical for all contaminants and shown exemplarily in this paper for nitrate.

As a second example, we focus on an area of heterogeneous land use (inset on Fig. 8c), where seven locations were selected. At those locations, the change of the mixed local distributions for different neighbourhood sizes is demonstrated (Fig. 8). With a neighbourhood size of zero, only three distribution functions exist, because the seven selected locations are in cells with three different land use groups (group 2 indicated by orange with points 3 and 7; group 0 indicated by red with points 1 and 2; and group 5 indicated by green with points 4, 5, and 6 in Fig. 8).

With a neighbourhood size of $nneib = 32$, location 5 has a fairly homogeneous neighbourhood, consisting mostly of forest and hence the mixed distribution is the one that skews most to small nitrate concentration values. Contrary, at location 7, the land use consists mostly of farm fields, leading to the statistically most right distribution. The distributions at the other locations match this scheme nicely, according to the composition of the land use in the vicinity. The larger the neighbourhood gets, the more it is influenced by other types of land use, hence the difference becomes again smaller. For example for a neighbourhood size of $nneib = 128$, the difference between the different locations becomes negligible (Fig. 8 bottom panel).

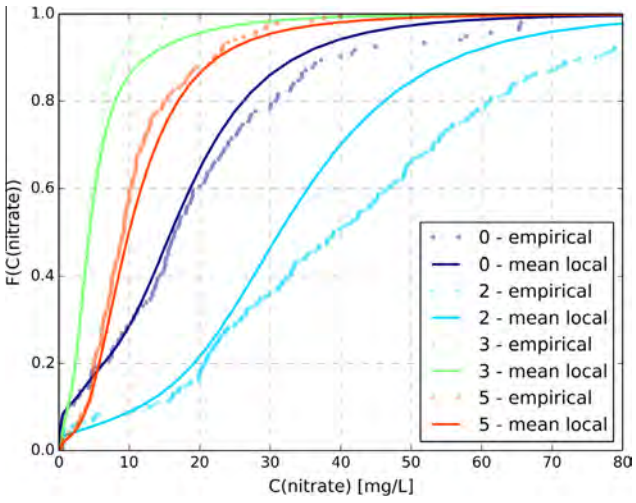

Fig. 5. Empirical nitrate distributions per land use group (here for groups 0, 2, 3 and 5) of measurements with relatively pure neighbourhoods (more than 70% of a single land use group), compared to mean mixed local distribution at identical locations for neighbourhood size of $nneib = 48$.

7. Optimal neighbourhood size

It has been established that the consideration of a certain neighbourhood has an effect on the locally mixed distributions. It remains to be investigated what the optimal neighbourhood size is. This question can be answered from three standpoints:

Cross-validation is the first way to determine an ideal neighbourhood size: The proposed method is able to estimate an expected concentration at an arbitrary location using a concentration that can be estimated at a locations where measurements exist. These estimates can then be cross-validated against the measurements. Leave-one-out cross-validation (Hawkins et al., 2003) was performed using different neighbourhood-sizes. Leave-k-out cross-validation was also performed, but did not lead to differing results. The measures of Mean Absolute Error (MAE, Eq. (13)), Root Mean Square Error (RMSE, Eq. (14)), and Linear Error in Probability Space (LEPS, Eq. (15)) were calculated and are presented in Table 4.

$$\mathrm{MAE} = 1/n \cdot \sum_{i=1}^{n} (|c_{mess}(i) - c_{interpol}(i)|) \tag{13}$$

$$\mathrm{RMSE} = \sqrt{1/n \cdot \sum_{i=1}^{n} (c_{mess}(i) - c_{interpol}(i))^2} \tag{14}$$

$$\mathrm{LEPS} = 1/n \cdot \sum_{i=1}^{n} (F[c_{mess}(i)] - F[c_{interpol}(i)]) \tag{15}$$

Method "A1" denotes the use of the expected value of the global distribution. Method "A2" denotes the proposed method of using the expected value of locally mixed distributions. Method A2 outperforms method A1. The inclusion of any neighbourhood outperforms method A1 and A2 with zero neighbourhood. Based on the three cross-validation methods, a neighbourhood size of ~ 2 km ($nneib = 64$) is best. It should be noted that for neighbourhoods larger than $nneib = 32$, little improvement is detectable and computational costs increase.

Change in distribution functions is the second way to determine an ideal neighbourhood size. Both pure distribution functions

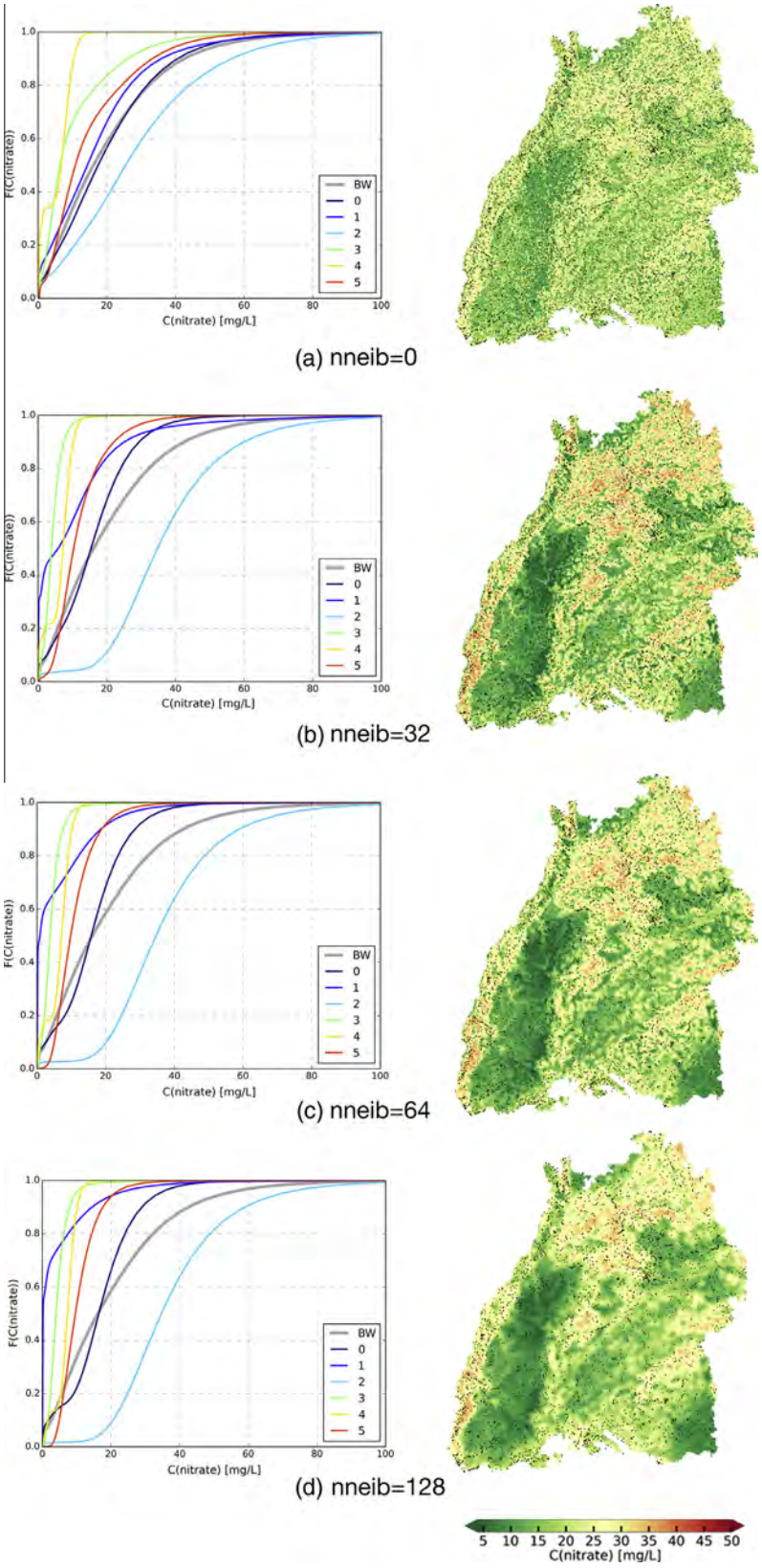

Fig. 6. Optimized "pure" distributions (left panels) and maps of the expected value of locally mixed distributions (right panels) for different neighbourhood sizes for nitrate in Baden-Württemberg. These maps do not take a model of spatial dependence into account.

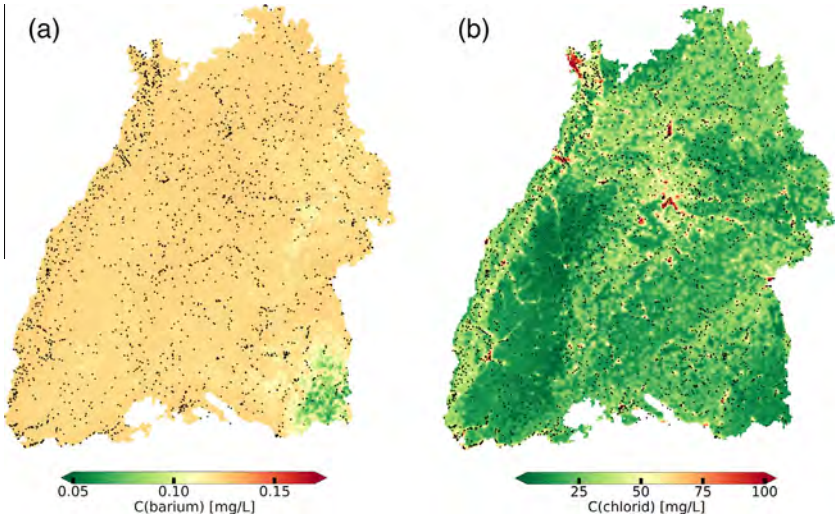

Fig. 7. Maps of the expected value of locally mixed distributions for barium (panel a) and chloride (panel b) for a neighbourhood radius of ~2 km. Fig. 6c shows the corresponding map for nitrate. The general structure is similar for nitrate and chloride, even though large chloride concentrations are more locally confined. Barium's concentration pattern is different compared to nitrate and chloride: land use has a much minor effect on barium.

(Fig. 6) and locally mixed distribution functions (Fig. 8) are suitable for this analysis. The global distribution functions change significantly from $nneib = 0$ to $nneib = 32$. From there on to larger neighbourhood sizes, the deviation from the global distribution of Baden-Württemberg remains visually very similar. The local distributions are based on seven locations only, but they show a similar picture as the pure distributions. When looking at a large neighbourhood size of $nneib = 128$, the heterogeneity of the neighbourhood resembles the heterogeneity of the domain more than for smaller neighbourhoods. Based on this analysis, a neighbourhood size between $nneib = 32$ and $nneib = 64$ seems ideal.

The difference between the global and the locally mixed distributions is the third possibility to optimize the ideal neighbourhood size. It can be calculated not only at locations where measurements are available (as done in cross-validation), but at regularly spaced locations over the domain. The mean of these deviations in probability space is a measure comparable to the LEPS (both are in probability space), and is listed in the column "Domain-LEPS" in Table 4.

8. Information content of secondary information

The information content (IC) at location j of the secondary information can be evaluated by calculating the maximum distance in probability space between the global and the locally mixed distribution function (Eq. (16)). The global distribution is identical at every location within the domain, the locally mixed distribution depends on the composition of the land use.

$$\mathrm{IC}_j = \max |F_{\mathrm{global}}(C_i) - F_{\mathrm{local}}(C_i)| \tag{16}$$

Maps of IC were calculated not only for nitrate, but also for the contaminants of chloride (that like nitrate is often of anthropogenic origin) and barium (generally of geogenic origin). The patterns of IC are similar for the two anthropogenic contaminants nitrate (Fig. 9a) and chloride (Fig. 9b). Nitrate exhibits the slightly larger gain of information from land use compared to chloride (shadings of darker reds exist). The IC of land use for barium shows a different pattern compared to the information content patterns of nitrate and chloride (Fig. 9c). For barium, the worth of information of land use is significantly decreased compared to the other two anthropogenic contaminants.

These maps are useful for a decision process if the selected type of secondary information is considered to influence the primary variable and should be included for spatial interpolation. Here, this is not the case for barium.

There is a lot more clustering of high nitrate values occurring than for low nitrate values. This is attributable to the fact that agricultural land is clustering more than forest. The differences between nitrate and chloride are worthy to be pointed out: large chloride values clump less and at different locations compared to large nitrate values. It could be argued that chloride contamination originates in different land use groups to nitrate, probably more in urban settings. It seems like these urban areas are more connected than the agricultural areas.

The IC can be averaged over the interpolation domain and serve as another performance measure for the spatial estimation method A2, and as a fourth measure to determine an ideal neighbourhood size (Table 4, in addition to the three measures presented in Section 7). The average information content over the domain (AICD, Eq. (17)) is less direct as the above used measures for cross-validation.

$$\mathrm{AICD} = \frac{1}{n_{ip\_loc}} \sum_{i=1}^{n_{ip\_loc}} \mathrm{IC}_i \tag{17}$$

The information gain from using secondary information gets larger as the neighbourhood considered increases in size. However, there is an optimum, as neighbourhood size goes to infinity (the domain), the usefulness of the secondary information decreases again. For the limit of a neighbourhood size of the domain, the resulting locally mixed distributions are the global distribution (see also Fig. 8). The optimum neighbourhood size based on AICD ($16 < nneib < 32$) is slightly smaller than for the cross-validation based measures ($32 < nneib < 64$).

9. Conclusions

Land use has a significant influence on groundwater quality parameters. This effect is even more pronounced if the composition of the land use in the vicinity of each estimation location is considered, and not only directly at the observation location.

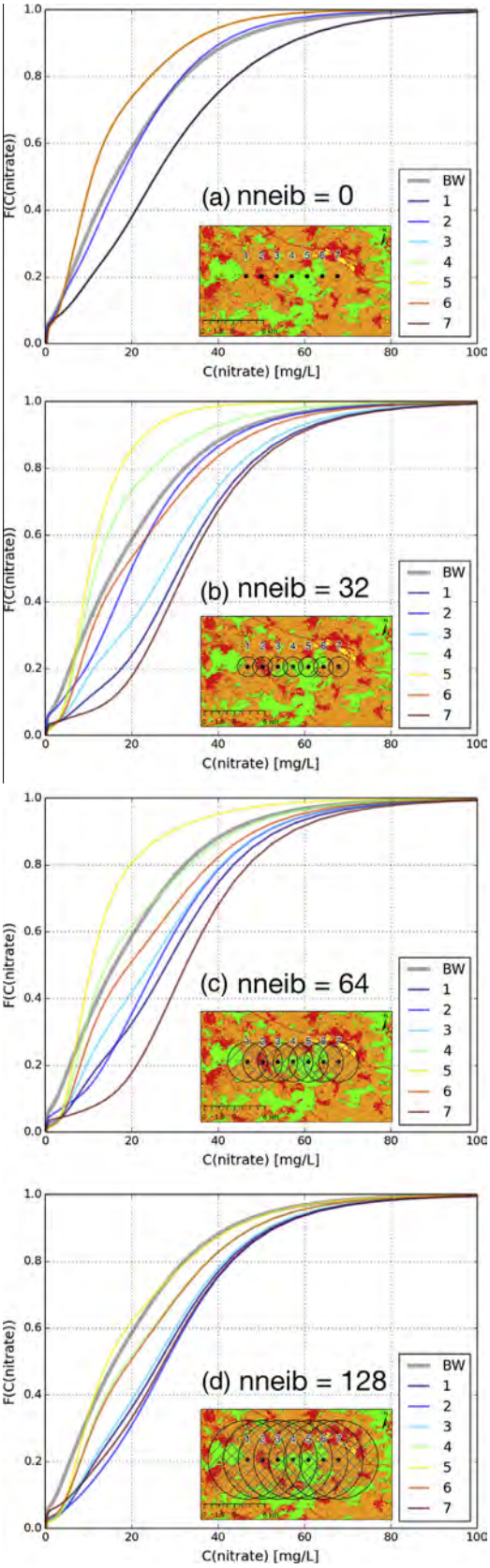

Fig. 8. Demonstration for mixing local distributions based on composition of the land use in the vicinity. The hatched areas indicate the neighbourhood whose size varies for each panel. The colours indicate the seven land use groups.

Table 4
Performance measures of local spatial averaging: cross-validation and average information content over the domain (AICD). The table compares taking the expected value of the global distribution (A1) with taking the expected value of locally mixed distributions (A2, different sizes of neighbourhoods) at locations where measurements exist. Mean average error (MAE) and root mean square error (RMSE) are in units of concentration; linear error in probability space (LEPS) and AICD are in probability space.

Method	*nneib*	Cross-validation based			Domain based
		MAE [C]	RMSE [C]	LEPS [–]	AICD [–]
A1	–	13.3	17.2	0.258	–
A2	0	12.1	16.3	0.231	0.111
A2	16	11.0	15.1	0.206	0.149
A2	32	10.6	14.8	0.197	0.149
A2	48	10.5	14.7	0.193	0.144
A2	64	10.4	14.7	0.192	0.140
A2	128	10.5	14.9	0.194	0.127

It is necessary to average land use spatially within a neighbourhood of a given size to derive an estimate of a concentration based on the land use in that neighbourhood. The method developed in this paper provides a local statement about the expected concentration based on the composition of the land use in the neighbourhood, which is an averaged measure. This local expected concentration can serve as additional information for spatial interpolation, which is analysed separately.

The pure distributions that are used for the mixing of the local distributions can be estimated with the procedures demonstrated in this paper. These pure distributions represent a homogeneous land use within a given neighbourhood size. The mixed distribution represents the actually occurring proportions of the land use groups within that neighbourhood. The mixed distribution at every estimation location is hence a mixture of the pure distributions of all the land use groups occurring within that neighbourhood.

By considering the land use composition of the neighbourhood only, about 25% of the variability could be explained by the locally mixed distributions compared to the global average. This method does not take a model for spatial dependence into account. Hence this represents a considerable improvement on previous methods.

The method reduces the danger of local estimation errors that originate from biased observation networks where observations are not uniformly spread over the categories of secondary information. Instead, in the demonstrated method the relative occurrence of categories of secondary information determines their weight for the final locally mixed distribution.

Depending on the contaminant, different kinds of secondary information could improve the statistical model. For example, land use categories can provide useful information for the spatial model of anthropogenic contaminants, whereas information on the spatial distribution of hydrogeological units can improve the interpolation of geogenic contaminants. Any kind of spatially distributed, readily available information can be incorporated in the presented approach, as the *local* information of the secondary information is incorporated via the marginal distribution.

Possibilities for improvement include (a) optimization of the weights of the groups of secondary information, in addition to the neighbourhood, and (b) match the shape of the neighbourhood included to the hydraulic or hydrogeologic conditions at the interpolation location. Differently shaped neighbourhoods could be optimal for areas with differing hydraulic or hydrogeologic characteristics. Other features, such as lines representing streams, could be included as an extra land use category to represent surface water/ground water exchange processes.

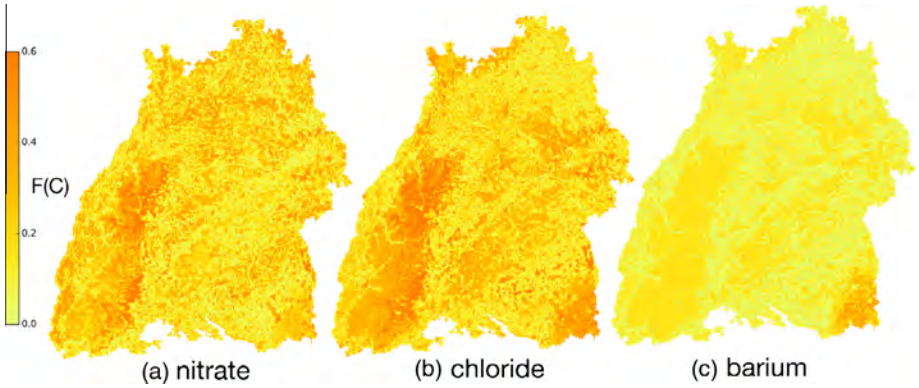

Fig. 9. Information content (IC, Eq. (16)) of locally mixed distributions (Eq. (6)) compared to the global distributions (Eq. (9)) for the contaminants nitrate (panel a), chloride (panel b), and barium (panel c) in the state of Baden-Württemberg.

Acknowledgements

Thanks to the state agency for the environment, measurement, and natural protection of the state of Baden-Württemberg (LUBW), Germany, for collaboration. This work was funded by German Research Foundation (DFG) grant GRK 1829/1 and DFG grant HA 7339/2-1.

References

Bronstert, A., Niehoff, D., Bürger, G., 2002. Effects of climate and land-use change on storm runoff generation: present knowledge and modelling capabilities. Hydrol. Process. 16, 509–529.

Cohen Jr., C.A., 1976. Progressively censored sampling in the three parameter log-normal distribution. Technometrics 18, 99–103.

Dueck, G., 1993. New optimization heuristics – the great deluge algorithm and the record-to-record travel. J. Comput. Phys. 104, 86–92.

Foley, J.A., DeFries, R., Asner, G.P., Barford, C., Bonan, G., Carpenter, S.R., Chapin, F.S., Michael T Coe, G.C.D., Gibbs, H.K., Helkowski, J.H., Holloway, T., Howard, E.A., Kucharik, C.J., Monfreda, C., Patz, J.A., Prentice, I.C., Ramankutty, N., Snyder, P.K., 2005. Global consequences of land use. Science 309, 570–574.

Foster, S., Cherlet, J., 2014. The Links Between Land Use and Groundwater. Electronic Source. <http://bit.ly/1yIM1kT> (access date: 27.08.15).

Haslauer, C.P., Rudolph, D.L., Thomson, N.R., 2005. Impacts of changing agricultural land-use practices on municipal groundwater quality: Woodstock, Ontario. Bringing Groundwater Quality Research to the Watershed Scale, vol. 29. IAHS Publication, pp. 223–233.

Hawkins, D.M., Basak, S.C., Mills, D., 2003. Assessing model fit by cross-validation. J. Chem. Inform. Model. 43, 579–586.

Helsel, D.R., Cohn, T.A., 1988. Estimation of descriptive statistics for multiply censored water quality data. Water Resour. Res. 23, 1997–2004.

Helsel, D.R., Gilliom, R.J., 1986. Estimation of distributional parameters for censored trace level water quality data: 2. Verification and applications. Water Resour. Res. 22, 147.

Jacobs, H., 2001. Abschlussbericht zur Erstellung einer Landnutzungskarte Baden-Württemberg auf der Basis von Satellitenbildern. Technical Report. Geo-Bild Ingenieurbüro. Karlsruhe.

Kroll, C.N., Stedinger, J.R., 1999. Development of regional regression relationships with censored data. Water Resour. Res. 35, 775–784.

Lerner, D.N., Harris, B., 2009. The relationship between land use and groundwater resources and quality. Land Use Policy 26, S265–S273.

Liu, S., Lu, J.C., Kolpin, D.W., Meeker, W.Q., 1997. Analysis of environmental data with censored observations. Environ. Sci. Technol. 31, 3358–3362.

Meiyappan, P., Jain, A.K., 2012. Three distinct global estimates of historical land-cover change and land-use conversions for over 200 years. Front. Earth Sci. 6, 122–139.

Mogheir, Y., de Lima, J.L.M.P., Singh, V.P., 2004. Characterizing the spatial variability of groundwater quality using the entropy theory: I. Synthetic data. Hydrol. Process. 18, 2165–2179.

O'Connell, P.E., Ewen, J., O'Donnell, G., Quinn, P., 2007. Is there a link between agricultural land-use management and flooding? Hydrol. Earth Syst. Sci. 11, 96–107.

Scanlon, B.R., Reedy, R.C., Stonestrom, D.A., Prudic, D.E., Dennehy, K.F., 2005. Impact of land use and land cover change on groundwater recharge and quality in the southwestern US. Glob. Change Biol. 11, 1577–1593.

Shumway, R., Azari, R., 2002. Statistical approaches to estimating mean water quality concentrations with detection limits. Environ. Sci. Technol. 36, 3345–3353.