



Available online at [www.sciencedirect.com](http://www.sciencedirect.com)

SCIENCE @ DIRECT®

ELSEVIER

Geoderma 130 (2006) 368–386

GEODERMA

[www.elsevier.com/locate/geoderma](http://www.elsevier.com/locate/geoderma)

## Accounting for change of support in spatial accuracy assessment of modelled soil mineral phosphorous concentration

Ulrich Leopold<sup>a,\*</sup>, Gerard B.M. Heuvelink<sup>b</sup>, Aaldrik Tiktak<sup>c</sup>,  
Peter A. Finke<sup>d</sup>, Oscar Schoumans<sup>e</sup>

<sup>a</sup>*Institute for Biodiversity and Ecosystem Dynamics, Universiteit van Amsterdam, Nieuwe Achtergracht 166, 1018 WV Amsterdam, The Netherlands*

<sup>b</sup>*Soil Science Centre, Wageningen University and Research Centre, P.O. Box 37, 6700 AA Wageningen, The Netherlands*

<sup>c</sup>*Netherlands Environmental Assessment Agency, National Institute for Public Health and Environment (RIVM), P.O. Box 1, 32720 BA Bilthoven, The Netherlands*

<sup>d</sup>*Department of Geology and Soil Science, University of Gent, Krijgslaan 281 S8, 9000 Gent, Belgium*

<sup>e</sup>*Centre for Water and Climate, Alterra, Wageningen University and Research Centre, P.O. Box 47, 6700 AA Wageningen, The Netherlands*

Received 10 August 2004; received in revised form 16 February 2005

Available online 29 April 2005

---

### Abstract

Agricultural activities in the Netherlands cause high nitrogen and phosphorous fluxes from soil to ground- and surface water. A model chain (STONE) has been developed to study and predict the magnitude of the resulting ground- and surface water pollution under different environmental conditions. STONE has three main components, namely: 1) the fertiliser distribution model CLEAN; 2) the atmospheric transport and deposition model OPS; and 3) the soil and soil-water quality model ANIMO. The goal of this study was to assess the accuracy of the STONE model output by comparing model predictions with independent observations, while accounting for spatial variation and differences in spatial support. The specific STONE output considered was mineral phosphorous concentration in the top soil. Predictions from STONE are made at the block support and, specifically, for 500 × 500 m grid cells, whereas the observations are on a point support. Therefore, aggregation of the point observations or disaggregation of the STONE model predictions is needed in order to test the accuracy of the STONE model. This study used the aggregation route. Specifically, ordinary and regression block kriging were used to aggregate the point observations to the block support. Overall, there was a good correspondence between the kriged observations and the STONE predictions, with no evidence of bias in the model predictions. However, there were meaningful differences locally, which was partly attributed to the smoothing of the kriging interpolator. Closer inspection revealed that not all of the differences could be attributed to measurement errors in the phosphorous observations and interpolation errors in the kriged phosphorous maps. Thus, it was demonstrated that the STONE output suffers from input and/or model errors. However, quantification of these errors was restricted by the fairly large interpolation uncertainty of the

---

\* Corresponding author. Tel.: +31 20 5257456; fax: +31 20 5257431.

E-mail addresses: [uleopold@science.uva.nl](mailto:uleopold@science.uva.nl) (U. Leopold), [Gerard.Heuvelink@wur.nl](mailto:Gerard.Heuvelink@wur.nl) (G.B.M. Heuvelink), [A.Tiktak@rivm.nl](mailto:A.Tiktak@rivm.nl) (A. Tiktak), [Peter.Finke@ugent.be](mailto:Peter.Finke@ugent.be) (P.A. Finke), [Oscar.Schoumans@wur.nl](mailto>Oscar.Schoumans@wur.nl) (O. Schoumans).

kriged phosphorous observations. Nevertheless, the spatial pattern of error in the STONE predictions could be partly attributed to combinations of landuse and soil type.

© 2005 Elsevier B.V. All rights reserved.

**Keywords:** Uncertainty; Regression-kriging; Geostatistics; Environmental process model; STONE

## 1. Introduction

Agricultural activities in the Netherlands, such as fertilising and manuring, cause high nitrogen and phosphorous inputs to the soil. These substances may leach to the ground- and surface water, causing environmental pollution. Legislation aims to limit agricultural pollution by restricting the amount and timing of manure and fertiliser applications. In order to develop such legislation, it is useful for policy makers to have access to (the results of) models that can explore the likely nature and consequences of agricultural pollution under different socio-economic or environmental scenarios. In recent years, much emphasis has been placed on the development of environmental process models that can predict future pollution under different agricultural scenarios on a national scale (Overbeek et al., 2001). In practice, however, the accuracy of these models predict may not be very high. The natural environment is a complex and heterogeneous system in which many non-linear processes operate. Regardless of how much effort is put into building and improving environmental process models, and to collecting the various inputs required by these models, it is generally accepted that their output will always be accompanied by non-trivial amounts of uncertainty (i.e. Loague and Green, 1991; Jansen, 1998). Clearly, given their important role in assisting policy making, it is important to quantify the uncertainties associated with environmental process models.

An uncertainty analysis begins with defining and characterising the sources of uncertainty in model predictions. Two main sources of uncertainty can be distinguished (Heuvelink, 1998; Jansen, 1998). These are input uncertainty, model uncertainty and uncertainty in observations. The main causes for input uncertainty are the generalisation, classification, interpolation, measurement and estimation errors. Model uncertainty arises because models are simplified

representations of reality, that purposely ignore less important processes or represent them in a crude way. Also, the numerical solution of a model often involves discretisations and approximations. Once defined and characterised, there are several techniques available to analyse how input and model uncertainties propagate to model predictions. The technique that is most often used for this is the Monte Carlo simulation (Loague and Corwin, 1996; Brown and Heuvelink, in press).

One of the main problems associated with performing an uncertainty analysis is the assessment of input and model uncertainties. Uncertainty information is rarely stored in databases, or is otherwise stored in a qualitative way, by means of summary-type meta-data (Shi et al., 2002). However, in principle it should be possible to assess and store the uncertainty in model inputs. For instance, measurement error can be derived from repeatedly measuring the same sample or may be provided by the instrument manufacturer. Interpolation error can be quantified by geostatistical techniques. In contrast, model uncertainties, are more difficult to assess because they are highly case-dependent. For example, a model that performs well under some conditions may perform badly when applied under other conditions, even when the quality of the input data is the same in both cases. Because model error is case-dependent, it can only be assessed realistically by comparing model predictions with independent observations using statistical techniques (Heuvelink and Pebesma, 1999). However, three important problems arise here, namely: 1) some model outputs cannot be measured or can only be measured indirectly or sparsely; 2) the support of model predictions is usually different from that of the observations; and 3) the differences between model predictions and observations are not only caused by model error but also by the propagation of input error and by measurement errors in the observations.

In this paper we address the problem of accuracy assessment of a soil process model (i.e., the STONE

model) that suffers from all three problems mentioned above (Fig. 1 left part). The overall objective of this study is to assess the accuracy of the model predictions by comparing them with independent observations. In doing so, we must account for differences in support between the observations and model predictions. Here, the observations are taken on a much smaller support ('point' support) than the aggregated model predictions which are taken on the 'block' support. We must also account for the presence of measurement errors in the observations. Moreover, as the observations are only available at a limited number of points, we must also consider interpolation uncertainty. The solution to this problem involves aggregation of the observations or disaggregation of the model predictions to achieve common support (Heuvelink and Pebesma, 1999; Bierkens et al., 2000). Thus, two possible routes can be followed: 1) interpolate and aggregate the observations to the model scale, i.e. in this case block support, and compare the 'aggregated observations' with the model predictions (Fig. 1 route with solid lines); (2) obtain the model predictions on a point support, compare these with the independent

observations, and aggregate the results of the comparison to the desired block support (Fig. 1 route with dashed lines). This study follows the first route.

The STONE model is used in the Netherlands to predict nitrate and phosphorous leaching from agricultural land to ground- and surface waters. In this study we assess the accuracy with which the STONE model predicts mineral phosphorous in the top 30 cm of soil, at the block support. Prediction and comparison on point support is considered unrealistic because of high uncertainties in the model inputs. Also, STONE is not intended to predict nitrate and phosphorous at points, but rather to predict the average nitrate and phosphorous concentration for larger blocks of land. Evaluation of the accuracy of STONE should, therefore, take place at a block support. Independent measurements for oxalate-extractable  $P$  are used to assess the accuracy of model predictions. Knowledge about the overall accuracy of the STONE model predictions is of key importance in understanding the value of the STONE model. It is also an important step towards the most important sources of uncertainty.

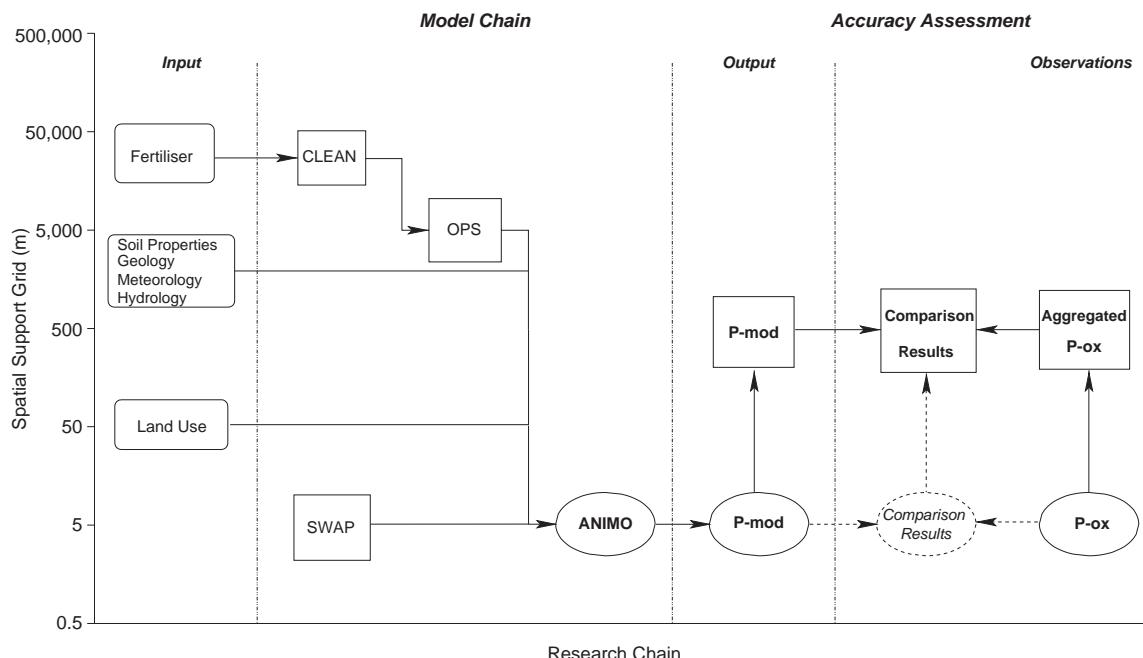


Fig. 1. Building blocks of STONE operate at different spatial supports. Comparison of STONE output (P-mod) and observations (P-ox) can be done on the block support (route with solid arrows) and point support (route with dashed arrows).

## 2. Data and methods

### 2.1. The STONE model

The STONE model was developed to study the effects of changes in agricultural practice and in policy measures on future soil as well as ground- and surface water pollution (Overbeek et al., 2001; Wolf et al., 2003). STONE concentrates on nitrogen and phosphorous pollution. In fact, STONE is a chain of three existing models (Fig. 1 left part). These are the fertiliser distribution model CLEAN, the atmospheric transport and deposition model OPS/SRM and the soil and soil water quality model ANIMO (Wolf et al., 2003). In addition, there are two supplementary models that are not used in calculations but provide inputs as boundary conditions to ANIMO. These are the hydrological model SWAP and the crop nutrient uptake model QUAD-MOD (Wolf et al., 2003). STONE is designed to make predictions for the whole of the Netherlands. Typically, the output is aggregated to a  $500 \times 500$  m block support or larger (Wolf et al., 2003).

Running STONE involves several steps. First, one or more policy measures or scenarios are defined that specify the manure production and amount of inorganic fertiliser use. Next, CLEAN computes the distribution of N and P input into the soil by manure and inorganic fertiliser application. CLEAN also computes ammonia emissions released from agriculture into the atmosphere. The OPS/SRM model is then applied to the ammonia emissions to calculate N deposition from the atmosphere to the soil. Next, ONAT and ANIMO are used to calculate carbon and nutrient cycling in soil and N and P fluxes to ground- and surface water bodies. N and P uptake by crops is calculated and provided by QUAD-MOD. The outputs of STONE include maps of N and P contents in soils, N and P concentrations in the shallow groundwater as well as fluxes of N and P into the groundwater and surface water.

The three main components of STONE each operate at their own spatial support because they address different processes and use inputs on different spatial supports. CLEAN requires input and generates output for 31 agricultural regions in the Netherlands, the so-called ‘LEI regions’. OPS/SRM computes deposition of N from the atmosphere for a  $5 \times 5$  km

grid over the Netherlands. GONAT/ANIMO requires input and generates output on the point support. Here, ‘point’ refers to an area of about  $5 \times 5$  m in size. Output is aggregated later to  $500 \times 500$  m block support.

Aggregation and disaggregation procedures have been developed to link the models used in STONE. CLEAN output needs to be disaggregated from LEI region support to the support of OPS/SRM. This is done on the basis of the spatial distribution of combinations of landuse type and soil type within each  $5 \times 5$  km grid cell. The OPS/SRM results need to be disaggregated to a  $500 \times 500$  m grid. This is done by assuming that the deposition within each  $5 \times 5$  km grid cell is constant.

#### 2.1.1. Main models in STONE

**2.1.1.1. CLEAN.** The main goals of CLEAN are to calculate the distribution of N and P inputs to agricultural soils and to calculate ammonia emissions from agricultural sources, (mainly manure) to the atmosphere. Total annual N and P inputs are available for the 31 LEI regions. Fertiliser use and manure production within these regions is estimated from landuse type and soil type. Hereby, six soil types and seven landuse types are distinguished. CLEAN results are derived through the following steps: 1) manure production of animal stock and excreted manure per animal category are calculated; 2) the resulting manure is distributed within a region by taking soil type/crop combinations into account; 3) surplus manure is allocated to other regions; 4) ammonia volatilisation and resulting N and P input into the soil are determined; and 5) application of inorganic fertiliser per region is calculated. The outputs are passed on to OPS/SRM and ONAT/ANIMO. Further information about CLEAN is given in Wolf et al. (2003).

**2.1.1.2. OPS/SRM.** The Source Receptor Matrix (SRM) model is applied within STONE as a simplification of the Operational Priority Substances (OPS) model. Both SRM and OPS calculate the mean annual N deposition from the atmosphere for a  $5 \times 5$  km grid over the Netherlands. SRM is a meta-model of OPS and was developed to reduce computing time, which is significant compared to other models in

STONE. The inputs of SRM and OPS are derived from the CLEAN output of ammonia emissions provided on a  $5 \times 5$  km grid. OPS calculates the N deposition levels using long term mean meteorological data. SRM uses these results to adjust the N depositions by assuming a linear relationship with ammonia emissions. More information about the modelling approach of OPS/SRM can be found in Asman and Van Jaarsveld (1992) and Wolf et al. (2003).

**2.1.1.3. GONAT/ANIMO.** Soil processes are calculated with the ONAT/ANIMO system. ONAT is an interface for collecting and preparing input data used by ANIMO. ONAT is also used for post-processing the ANIMO results, e.g. aggregation of outputs per region. ANIMO is a complex, process-oriented model for simulating carbon and nutrient cycling in soils. The most important components of ANIMO are: 1) a soil chemical transport module; 2) an organic carbon module; 3) a nitrogen cycle module; and 4) a phosphorous cycle module. Inputs to ANIMO are obtained in various ways. The manure and fertiliser applications are supplied by CLEAN and the N deposition by OPS/SRM. The hydrological data, such as soil water fluxes, soil water contents, drainage fluxes and evapo-transpiration fluxes are derived with SWAP (Kroes et al., 2000 and Van Dam, 2000). Basic soil properties are obtained by combining a Soil Information System with the 1:50,000 digital soil map of the Netherlands (Kroon et al. (2001)). Specific soil properties required by ANIMO are derived directly from measurements or indirectly using pedo-transfer functions.

In this article we focus on the phosphorous cycle in ANIMO. The decomposition and transformation processes of fresh organic materials and the humus turnover determine the mineralisation and immobilisation of P in soils. In the organic phosphorous cycle the following processes are described: 1) application of various organic materials, such as manure and crop residues; 2) decomposition of roots and root exudate; 3) decomposition of fresh organic materials in soils and transformation to humus and 4) turnover of humus.

In the inorganic part of the P cycle the following processes are described: 1) addition of mineral P in organic and inorganic fertilisers; 2) P sorption; 3) P

leaching and 4) P uptake by crop. P uptake by crop is calculated in QUAD-MOD, which assumes a linear relationship between nutrient application and nutrient uptake and a curvi-linear relationship between biomass yield and nutrient uptake. Further details can be found in Schoumans and Groenendijk (2000) and Wolf et al. (2003). ANIMO calculates vertical transport of P through a one-dimensional system (i.e. the unsaturated zone of the soil). In addition, lateral fluxes of P to various drainage systems and surface water (e.g. ditches, canals) and vertical fluxes to deep soil layers and groundwater can be incorporated.

## 2.2. Obtaining STONE output at the target support

### 2.2.1. Running STONE for the whole of the Netherlands

STONE must be run at the point support because it solves partial differential equations that are only valid at the point support (Heuvelink and Pebesma, 1999). However, running a complex model such as STONE at the nodes of a high resolution grid for the Netherlands would require too much computation time. Therefore, each grid node is assigned to one of a limited number of classes or so-called unique units, which are formed by combinations of basic soil properties, land-use types, climate districts and geo-hydrological districts. The classes are defined such that the (rounded) inputs to all nodes in the class are the same. As such, STONE only needs to be run once for all nodes occupying the same class.

The input data for STONE come from different sources. Groundwater depth classes are derived from the 1:50,000 digitised soil map of the Netherlands. Seepage fluxes are based on calculations from a regional groundwater model (De Lange, 1999). Drainage characteristics are calculated analytically, with some simplifying assumptions (De Lange, 1999). Landuse types are based on satellite imagery, whereby only four major land uses are distinguished, namely permanent grassland, maize, other arable land and nature. The variability of weather conditions across the country is represented by 15 climate districts, as published by the Royal Dutch Meteorological Institute. Soil profiles and soil chemical properties are derived from the national soils database (Kroon et al., 2001). Within STONE, a total number of 456 soil profiles are distinguished.

All original maps are converted to raster maps with a resolution of 250 m. The raster maps are combined into a map that has more than 100,000 different input combinations. This number was reduced to 6405 units by grouping similar inputs into the unique units mentioned above. This is done using so-called relation diagrams (Kroon et al., 2001), which merge analogous properties into the same class.

### 2.2.2. Aggregating STONE output to the target support

In order to compare the STONE output with the phosphorous observations it is necessary to define a spatial support on which the comparison can be made. On the national level, it is neither realistic nor desirable to assess how well STONE can predict phosphorous concentrations at points. Rather, the interest is in the average soil phosphorous concentration for larger areas. Here, the size of the larger area is chosen as blocks of 500 × 500 m support. This provides sufficient resolution to detect spatial patterns while ensuring that the accuracy of the STONE output is not evaluated at supports for which it was not intended.

Using a 500 m grid size results in 112,795 cells for the whole of the Netherlands (urban areas are excluded). In addition, we took the 0–30 cm top soil for comparison of STONE results with observations. This part of the soil is heavily influenced by agricultural practice. The temporal target support, which is an average over the years 1993–2000, was determined by the sampling period for the P-ox observations (see Section 2.3).

STONE was run with scenario MV5 (Tiktak et al., 2003). Simulation results were obtained as time series of 10 day averages over the entire period from 1993 to 2000, for all 6405 unique units and for all 22 depth layers that STONE considers. The following steps were necessary to derive P-mod at the target support: 1) P-mod predictions in kg/ha were transformed to mmol/kg units, using bulk density information; 2) for each unique unit and each 10 day period, the average P-mod concentration of the top 30 cm was extracted from the P-mod concentrations of the entire vertical profile; 3) the 10 day averages were aggregated to an average for the whole 1993–2000 period and spatially aggregated to the 500 × 500 m block support.

### 2.3. Aggregating point observations to the target support

In order to assess the accuracy of the STONE output we used an independent data-set, the so-called LSK data-set. The LSK data-set has been obtained from a Dutch national sampling campaign carried out from 1993 to 2000. The purpose of this campaign was to upgrade the 1:50,000 soil map of the Netherlands by developing a database containing a statistically representative sample of Dutch soils (Finke et al., 2002). Details of the LSK data-set are presented below.

#### 2.3.1. Sampling

To obtain a statistically useful as well as cost effective data set a stratified random sampling design (Brus and de Gruijter, 1997) was employed. The Netherlands was divided into nine primary strata on the basis of groundwater table regions. Next the primary strata were subdivided into 77 secondary strata on the basis of soil and landuse maps (Finke et al., 2002). Within these stratified areas sampling locations were chosen at random. The total number of sampling locations is 1387. Bulk samples were taken by collecting subsamples in a 5 m<sup>2</sup> area. Soil samples were taken from the surface to the mean highest water table depth. This resulted in a set of soil samples with depths ranging from 15 to 280 cm. To characterise the soluble fraction of phosphorous for the complete soil profile, the concentration [mmol/kg] of oxalate-extractable P (P-ox) was measured for each soil horizon. A detailed description of the extraction method is given in Koopmans et al. (2001). Furthermore, various soil properties, including pH, bulk density, Al-ox, Fe-ox, soil texture, soil type, as well as landuse type were measured or determined for each sampling location.

#### 2.3.2. Pre-processing

Some pre-processing on the data collected was necessary for this study. In order to obtain oxalate-extractable P concentrations for a soil depth of 0–30 cm we computed the weighted average by taking the fraction of horizons within the 30 cm into account.

#### 2.3.3. Geostatistical aggregation

As we followed the aggregation route in Fig. 1 (solid arrows) we needed to aggregate the point

observations of P-ox to spatial blocks of  $500 \times 500$  m. For this we used ordinary and regression block kriging (Goovaerts, 1997; Hengl et al., 2004) with a discretisation grid of 16 nodes. Kriging assumes that the variable of interest, P-ox in our case, may be interpreted as a realisation of a random function  $Z$  whose statistical properties are assumed to be known (Chiles and Delfiner, 1999; Goovaerts, 1997). Ordinary kriging uses weighted linear combinations of the observations of  $Z$  to predict  $Z$  at unobserved locations. The weights are chosen such that the prediction error variance is minimised, under the condition of unbiasedness. The weights are a solution of the kriging system, which can be expressed in terms of semi-variogram values, which describes the spatial autocorrelation structure of  $Z$ . In mathematical terms, the ordinary kriging predictor is defined as

$$\hat{Z}(B) = \sum_{i=1}^n \lambda_i Z(x_i) \quad (1)$$

where  $\hat{Z}(B)$  is the prediction of the average value of  $Z$  over a block  $B$ , obtained from the  $n$  observations  $Z(x_i)$ , using weights  $\lambda_i$ .

Regression kriging is an extension of ordinary kriging. Whereas ordinary kriging assumes that the mean of  $Z$  is constant, in regression kriging the mean is taken as a non-constant trend, defined as a linear regression on explanatory variables. Thus, in regression kriging the random function  $Z$  is treated as the sum of a deterministic trend  $m$  and a zero-mean stochastic residual  $\varepsilon$ :

$$\hat{Z}(x) = m(x) + \varepsilon(x) = \sum_{k=0}^m \beta_k y_k(x) + \varepsilon(x) \quad (2)$$

where the  $\beta_k$  are regression coefficients and the  $y_k$  are explanatory variables (usually  $y_0$  is taken as identical to 1 so that  $\beta_0$  is the intercept of the regression). Prediction using regression block kriging is done using:

$$\hat{Z}(B) = \sum_{k=1}^m \hat{\beta}_k y(B) + \sum_{i=1}^n \kappa_i \varepsilon(x_i) \quad (3)$$

The regression coefficients are estimated from the observations using weighted least squares. Note, that we can use here the regression coefficients derived at the point support for estimations on the block support as the regression model is linear and, thus, the block

predictions and the corresponding prediction variance are not affected by a change of support (Heuvelink and Pebesma, 1999). The kriging weights  $\kappa_i$  are solved by the kriging system, which can be expressed as semi-variogram values of the residuals. The corresponding regression kriging variance incorporates the uncertainty about the estimation of the regression coefficients (Hengl et al., 2004; Knotters et al., 1995). Some authors use the terms ‘kriging with external drift’ (e.g. Goovaerts, 1997; Wackernagel, 2003) or ‘universal kriging’ (e.g. Chiles and Delfiner, 1999; Christensen, 1990) instead of regression kriging.

Both ordinary block kriging (OBK) and regression block kriging (RBK) were used in this study. For OBK, first an experimental variogram was computed on the pre-processed P-ox values. Next, the experimental variogram was modelled by choosing a variogram shape and fitting the variogram parameters nugget, range and sill. An automatic fitting procedure was applied using ordinary least squares within the geostatistical software package Gstat (Pebesma and Wesseling, 1998; Pebesma, 2001). Finally, the P-ox concentrations were predicted and aggregated using Eq. (1).

For RBK of P-ox, candidate explanatory variables were various soil properties, soil type and landuse (see Table 1). The soil and landuse maps are obtained by reclassification of the original soil and landuse maps, which have more classes. This was done to ensure that a sufficient number of P-ox observations are available per class, to achieve reliable estimates of the regression coefficients. Soil type and landuse are measured on a categorical scale and were transformed to indicator variables to be used in Eq. (3) (Hengl et al., 2004). The soil properties were derived by pedo-transfer functions from the national 1:25,000 soil map. Explanatory variables were selected using stepwise backward

Table 1  
Additional information used in regression kriging

Information	Variables
Soil properties	pH, SOM, clay and loam content, media sand grain diameter
Soil type	Mixed, old clay, loamy soils, sea clay, young river clays, stony soils, peat soils, sandy soils
Landuse	Grass land, arable land, forest, green urban areas, other natural vegetation (e.g. dunes, bogs), other areas (e.g. industrial sites)

linear regression, using the Akaike Information Criterion (AIC). The regression analysis was carried out with the statistical software package R ([R Development Core Team, 2003](#)). Residuals at observation locations were computed by subtracting the fitted regression line from the observations. From the residuals an experimental variogram was computed and a variogram model was fitted. Finally, the P-ox concentrations were interpolated and aggregated using Eq. (3).

#### 2.4. Comparison of STONE output and aggregated observations

To assess the accuracy of the STONE output, a comparison was made with aggregated P-ox observations. For this, we subtracted the kriged prediction map, based on the LSK observations, from the STONE output map:

$$\Delta Z = \hat{Z}(\text{STONE}) - \hat{Z}(B) \quad (4)$$

where  $\hat{Z}(\text{STONE})$  represents the STONE output on  $500 \times 500$  grid cells and where  $\hat{Z}(B)$  refers to the block-kriged LSK observations of P-ox. This was performed for ordinary block kriging as well as for regression block kriging.

To determine whether the observed differences were statistically significant we took the kriging interpolation error into account. Accordingly, the differences  $\Delta Z$  were standardised as follows:

$$\Delta Z_{\text{std}} = \frac{\Delta Z}{\sigma_{\text{krig}}} \quad (5)$$

with  $\sigma_{\text{krig}}$  being the block kriging standard deviation. Hence, we are able to determine whether the observed differences may be due to interpolation uncertainties or whether part of the observed differences can be attributed to a deviation of the STONE output from the true block-averaged P-ox concentration. We considered the latter to be the case when the standardised difference computed with Eq. (5) is greater than +2 or smaller than -2.

To determine the magnitude of the differences as a fraction of the STONE predictions, the differences were normalised as follows:

$$\Delta Z_{\text{norm}} = \frac{\Delta Z}{\hat{Z}_{\text{STONE}}} \cdot 100 \quad (6)$$

The normalised differences were only computed for those cases where the difference was significant.

The resulting maps of differences and normalised differences were computed and their spatial pattern interpreted. Furthermore, the significant differences were plotted against the STONE predictions. Relationships with soil type and landuse were investigated, in an attempt to explain observed errors in the STONE output.

### 3. Results and discussion

#### 3.1. STONE output

[Fig. 2\(a\)](#) shows the P-mod concentration of the top 30 cm of the soil on a  $500 \times 500$  m block support for the whole of the Netherlands as predicted by STONE. Predicted P-mod concentrations are generally high in the peat soils of the central part of the country and low in the clay soils of the western and northern part of the country. In the sandy soils of the eastern and southern part of the country, the predicted P-mod concentrations are intermediate.

The large area fraction of low to medium P-mod concentrations and small areal fraction with high P-mod concentrations allude to positive skewness of the STONE model predictions. Indeed, 90 mmol/kg P-mod concentration in the topsoil is relatively high because the median concentrations are around 20 mmol/kg. Among others, STONE uses available soil and landuse information to compute P-mod concentrations. The spatial patterns seem to be mainly determined by soil and landuse information. High concentrations appear to coincide with intensive arable land and grassland on soils with high clay or organic matter content. These relationships can be seen by comparing the STONE map with the soil and landuse maps in [Fig. 4](#). Low P-mod concentrations occur where there is natural vegetation on sandy soils, such as the forests at the centre of the Netherlands and the sand dunes along the west coast.

#### 3.2. Aggregation of point observations

##### 3.2.1. Exploratory data analysis

The 1387 sampling locations are shown in [Fig. 2\(b\)](#). They are fairly evenly distributed over the

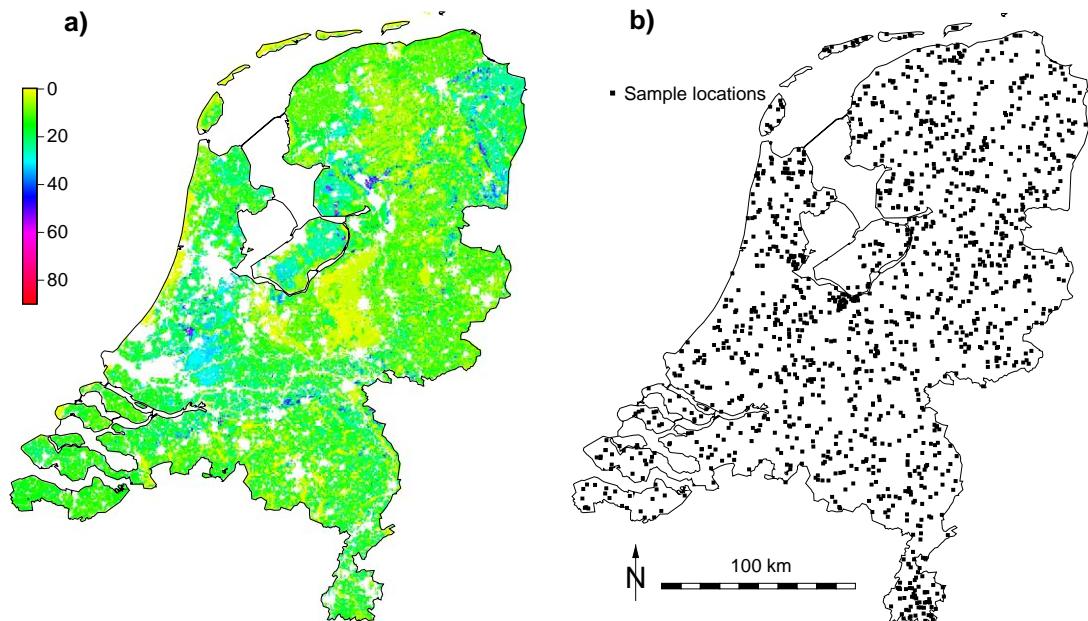


Fig. 2. STONE predictions for P-mod [mmol/kg soil] of top soil (left) and sampling locations for observations (right).

Netherlands, as a result of the stratified random sampling scheme. The histogram of the P-ox observations is shown in Fig. 3(a). The distribution of the data is positively skewed. The distribution shows some outliers between 50 and 100 mmol/kg and some extreme values over 100 mmol/kg. Log transformation of these data yields Fig. 3(b).

Transformation of data prior to the geostatistical interpolation and back-transformation afterwards is recommended when the data are strongly skewed. However, back-transformation is much more difficult if the analysis also includes an aggregation step (Chiles and Delfiner, 1999; Cressie, 1993). Here we decided to work directly with the untransformed data.

Analytical methods for back-transformation of block-averaged predictions and their corresponding variances do not exist and so the only viable option would involve a computationally demanding stochastic simulation approach (Chiles and Delfiner, 1999; Heuvelink and Pebesma, 1999). This was considered beyond the scope of the present study and we decided to apply kriging to the original data. The outliers and extreme values were not removed from the data set as we considered these values realistic.

High peaks can occur in regions with point sources such as former and current industrial sites or garbage dumps. It is also possible that a net accumulation of P-ox occurs where ground- and surface water flows converge.

Table 2 lists summary statistics for the stepwise linear regression. The variables chosen were clay content, median sand grain diameter (M50), SOM, peat soil, sandy soil, arable landuse, grassland and a group of various landuse types mainly consisting of industry. The minimum significance level chosen as an entry for the stepwise regression analysis was 0.05. The  $R^2$  appears to be low with a value of only 0.19. Fig. 3(c) shows the histogram of the residuals for the linear regression. The residuals are much more symmetrically distributed than the original observations. Apparently, part of the asymmetry in the P-ox observations is explained by the explanatory variables in the regression model. However, outliers and extreme values remain.

Fig. 4(a-d) show the spatial distribution of the explanatory variables selected for regression, namely soil type (a), landuse (b), clay content (c) and soil organic matter (SOM) (d). The eastern and southern parts of the country are characterised by sandy soils,

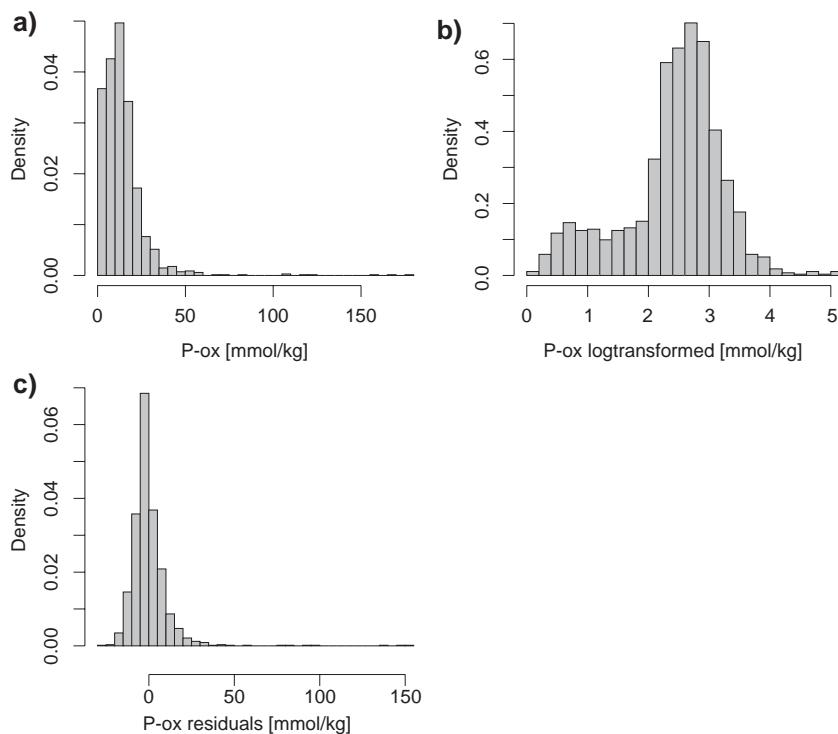


Fig. 3. Histograms for a) original observations, b) log-transformed observations and c) residuals.

whereas clay soils are mainly situated in the northern and south-western parts. Peat soils are found in the northern and central parts of the country. The clay and SOM maps are consistent with the soil map, but reveal more detail as these were based on a more detailed soil map. Dominant types of landuse are grassland and arable land, with forests and other natural vegetation occupying only

small parts of the Netherlands. Green urban areas are rather small and scattered.

The variables chosen by the stepwise linear regression are those we had expected to be correlated with P-ox concentrations. Soil clay content and SOM are responsible for bondage of P-ox in soil. Peat soils reflect high SOM contents, whereas in sandy soils low P-ox concentrations should be expected. The significance of arable land and grassland are consistent with our expectations that P-ox and phosphorous in general accumulate in soils with intensive agricultural landuse.

The low  $R^2$  can be partly explained by highly generalised soil information. Although landuse was available on a  $500 \times 500$  m resolution, this still represents a coarse spatial resolution for the purposes of estimating P-ox. Spatial variation of landuse within a  $500 \times 500$  m area can be substantial in the Netherlands, because landuse is often highly fragmented. Thus, we would not expect a strong relationship between the dominant landuse in a  $500 \times 500$  m grid cell and the P-ox concentration

Table 2  
Summary statistics for the linear stepwise regression

	Estimate	Std. error	t value	Pr (> t )	Signif.
(Intercept)	0.4960	1.4079	0.35	0.7247	
Clay	0.1851	0.0303	6.11	0.0000	***
M50	-0.0179	0.0072	-2.49	0.0130	*
SOM	0.1407	0.0411	3.42	0.0006	***
Peat soils	4.3751	1.4573	3.00	0.0027	**
Sandy soils	5.7717	1.1491	5.02	0.0000	***
Arable landuse	8.7863	1.1731	7.49	0.0000	***
Grassland	8.0917	1.0309	7.85	0.0000	***
Other areas	6.0440	2.5381	2.38	0.0174	*

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘.’ 1.

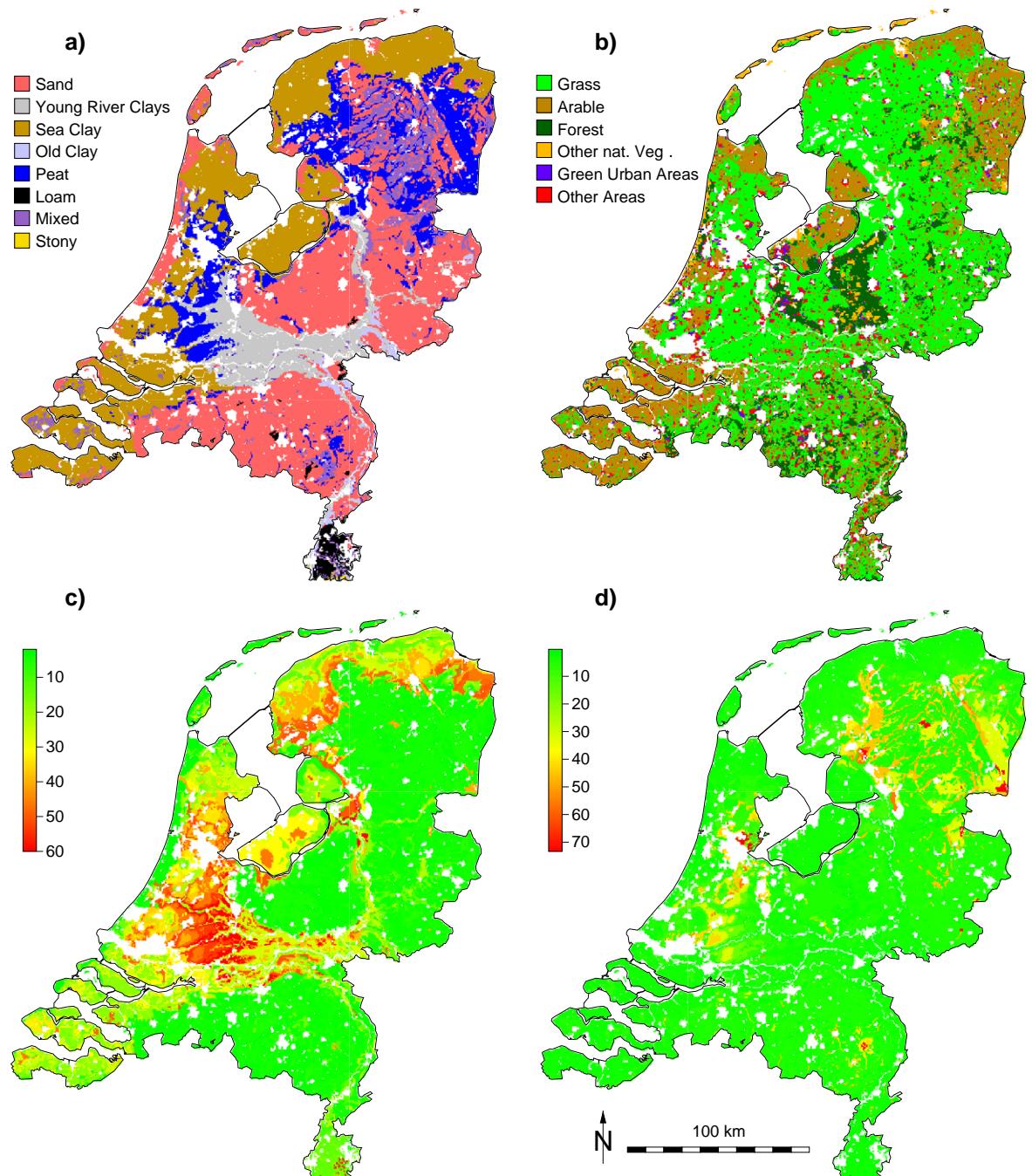


Fig. 4. Maps for explanatory variables used in regression block kriging of P-ox. Soil types (a), landuse (b), clay content [%] (c) and SOM contents [%] (d).

for a point location within the grid cell. Another reason for the poor  $R^2$  may be the non-linear relationship between P-ox and explanatory variables. Here we used linear regression to avoid complex problems of estimating regression parameters and performing a change of support, but the skewed distribution of P-ox does suggest that a non-linear regression may yield a larger  $R^2$ .

### 3.2.2. Geostatistical aggregation and prediction

The experimental variograms and the fitted variogram model for the ordinary and regression kriging are shown in Fig. 5(a) and (b), respectively. The nugget in the variogram for the untransformed observations is relatively high. It accounts for more than half of the sill for ordinary kriging (Fig. 5a). The range is approximately 40 km. For the variogram of the residuals in Fig. 5(b), all three parameters nugget, range and sill are smaller than the corresponding

parameters of the variogram of the original data. The nugget and sill decrease by about 20%. The range decreases from 40 to 25 km.

As the regression resulted in a low  $R^2$  of 0.19 we would expect only a slight improvement in regression block kriging over ordinary block kriging. This was indeed the case. With only 19% of the variance in P-ox explained by the additional information, the sill of the variogram of the regression residual was only about 19% smaller than that of the original data. Since the explanatory variables are smoothly varying quantities, the differences between the variograms in Fig. 5 are mainly reflected by a decrease in the sill and range, whereas the nugget variance remains largely unchanged. Goovaerts (1997) and Hengl et al. (2004) observed similar effects of regression on variogram parameters.

Fig. 6 shows the resulting kriging maps. The differences between OBK (Fig. 6a) and RBK (Fig. 6b) are small. Both maps show larger areas of low P-ox concentration in the north, centre and south of the country, as well as along the west coast. Areas with higher P-ox concentrations appear in the western and northern part of the Netherlands. The regression kriging map gives more detailed information about the spatial distribution of P-ox. The corresponding kriging variances are shown in Fig. 6(c) and (d). In general, the kriging variance is quite high both for OBK and RBK across large parts of the Netherlands. An interesting feature is that the regression kriging variance is larger than the ordinary kriging variance in most parts of the country.

Fig. 6(a) and (b) conform the observed differences in the variograms. The much smoother surface in the ordinary block kriging map reflects the larger variogram range. The areas with high P-ox concentrations can be explained by landuse and soil type. These are areas with intensive agricultural landuse or large industrial sites, where the soils have high clay and/or SOM content, to which P-ox is readily bound. Low P-ox concentrations occur in areas with natural vegetation on poor sandy soils.

Regression kriging improves the predictions of P-ox over ordinary kriging in terms of spatial detail (Bishop and McBratney, 2001; Hengl et al., 2004), which can be attributed to the additional variability of the explanatory variables. Nevertheless, the uncertainty in the regression coefficients resulted in a larger

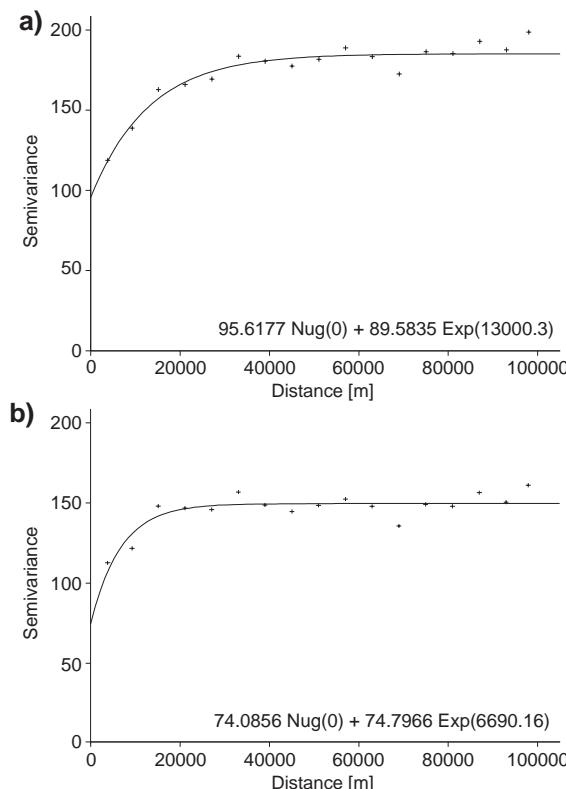


Fig. 5. Experimental semi-variogram (dots) and fitted model (solid line) of P-ox observations (a) and P-ox residuals (b).

overall interpolation uncertainty for the regression block kriging than for ordinary block kriging. This has been observed by others (Bishop and McBratney,

2001; Hengl et al., 2004). The spatial patterns in the kriging variance maps reflect the sampling design of the 1387 sample locations.

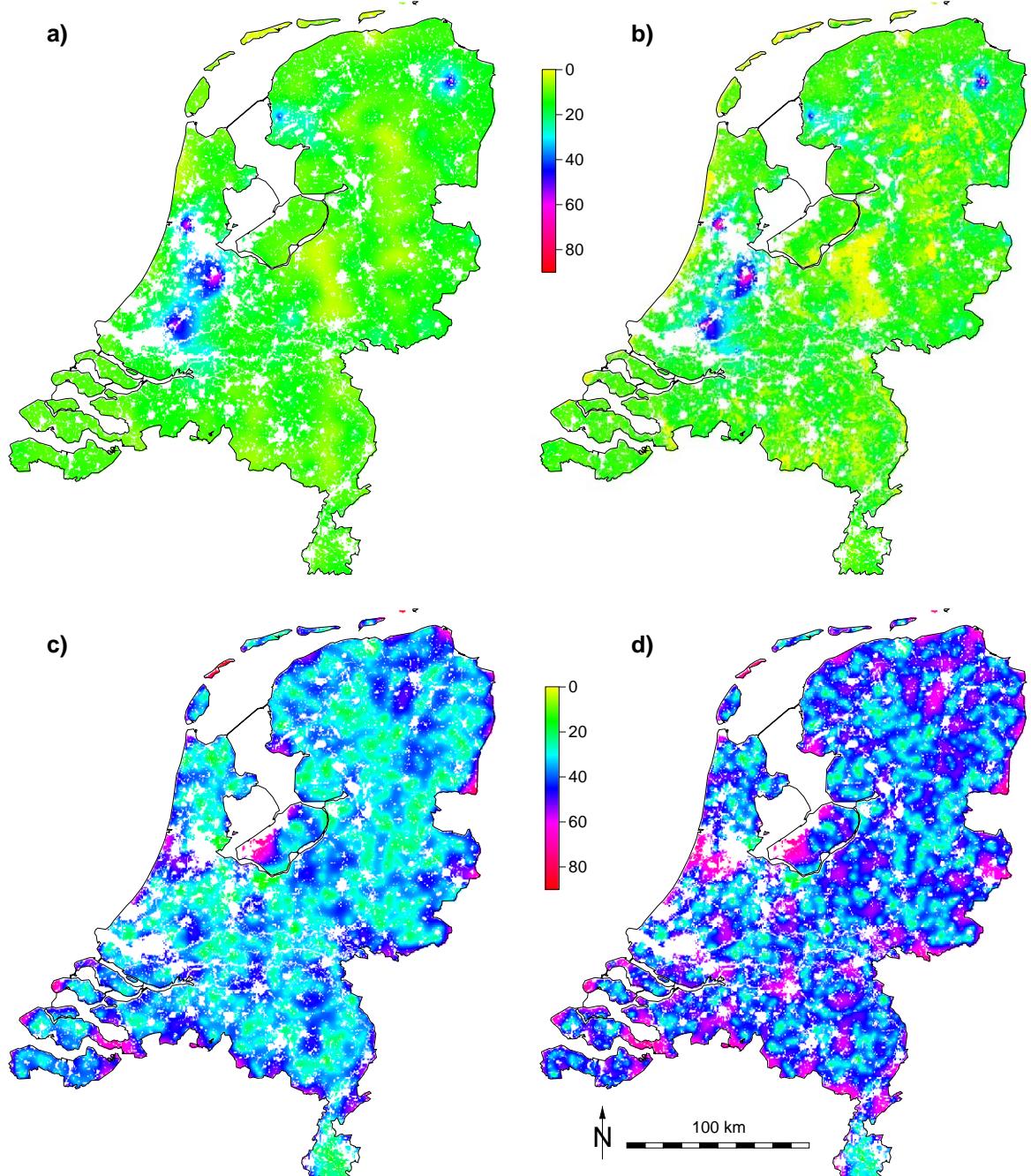


Fig. 6. Kriging predictions for P-ox [mmol/kg soil] and kriging variances [ $\text{mmol/kg soil}$ ] $^2$  of topsoil. a) ordinary block kriging, b) regression block kriging, c) ordinary block kriging variance and d) regression block kriging variance.

### 3.3. Comparison of STONE model output and aggregated observations

**Table 3** gives the global summary statistics for the STONE model, OBK and RBK predictions. The overall mean predictions are very similar whereas the variances show clear differences between the three predictors. Kriging involves more smoothing than STONE, and ordinary kriging even more so than regression kriging.

**Fig. 7(a)** and (b) show the local differences between the STONE and kriging predictions, for the ordinary and regression kriging cases, respectively. The right column shows residuals for STONE and OBK. The left column shows the residuals for STONE and RBK. There are no big differences between **Fig. 7(a)** and (b). Large areas of high and low values occur as well as scattered patterns in the south-east and north-east.

Clearly, there are similarities between the global patterns of the OBK and RBK predictions and those of the STONE predictions. There is an overall good correspondence between STONE and the kriged observations for the global mean. The average absolute difference for the STONE OBK comparison is 0.01 [mmol/kg soil] and for the RBK comparison 0.31 [mmol/kg soil]. However, **Fig. 7(a)** and (b) also reveal meaningful differences between the STONE predictions and kriged observations. These differences appear to be large in some parts of the Netherlands. In some areas, the differences for the RBK comparison are smaller than for the OBK comparison. Correlation coefficients for STONE and OBK and RBK are 0.34 and 0.50, respectively. [Kros et al. \(2002\)](#) obtained similar results.

**Fig. 7(c)** and (d) show only areas where significant errors occur. Recall that differences are considered significant when the standardised difference is larger than +2 or smaller than -2 (see Eq.

Table 3

Summary statistics for phosphorous concentration predicted with STONE, ordinary block kriging (OBK) and regression block kriging (RBK)

Predictor	Mean [mmol/kg soil]	Variance [mmol/kg soil] <sup>2</sup>
STONE	13.69	70.93
OBK	13.67	39.43
RBK	13.38	50.30

(5)). **Fig. 7(c)** and (d) show a clear reduction of the size of the total area where meaningful differences occur. For large parts of the Netherlands the differences are not significant. The comparison of STONE output with regression kriging predictions in **Fig. 7(d)** shows a smaller area with significant differences than the comparison with ordinary kriging in **Fig. 7(c)**.

If we take interpolation uncertainty into account, differences between STONE and either OBK or RBK are significant over a smaller part of the total area. Apparently, the interpolation uncertainty is sufficiently large to prevent attribution of the observed differences for large parts of the Netherlands. Comparison with RBK predictions yields fewer significant areas than comparison with OBK. Thus, either the interpolation error is larger for RBK than for OBK or the difference between the STONE predictions and aggregated observations is smaller for RBK. Closer inspection of the results shows that both phenomena occur. The smaller differences between STONE and RBK predictions may be explained by the reliance of both models on the same explanatory information (i.e. soil type and landuse). However, the smaller region of statistically significant differences in the RBK model is explained by larger interpolation uncertainty for RBK than OBK.

Closer inspection of the spatial distribution of significant differences reveals that STONE underpredicts mainly in those areas (blue) where grassland occurs on peat soils with high clay and SOM contents. The most likely explanation is that the high natural background values of phosphorous are underpredicted in these areas. However, underprediction may also be caused by underestimation of upward seepage of phosphorous rich water by STONE.

Overprediction (red) of P-mod by STONE occurs mainly in arable land on sea clay and peat soils. A possible reason for this is that transport of manure from farms with a manure surplus is underpredicted by CLEAN. Alternatively, it might reflect an incorrect assumption about the capability of some, particularly young, soils to bond phosphorous.

Apart from these two distinct areas of over- and underprediction, there are scattered areas of over- and underprediction in sandy soils (mainly the east

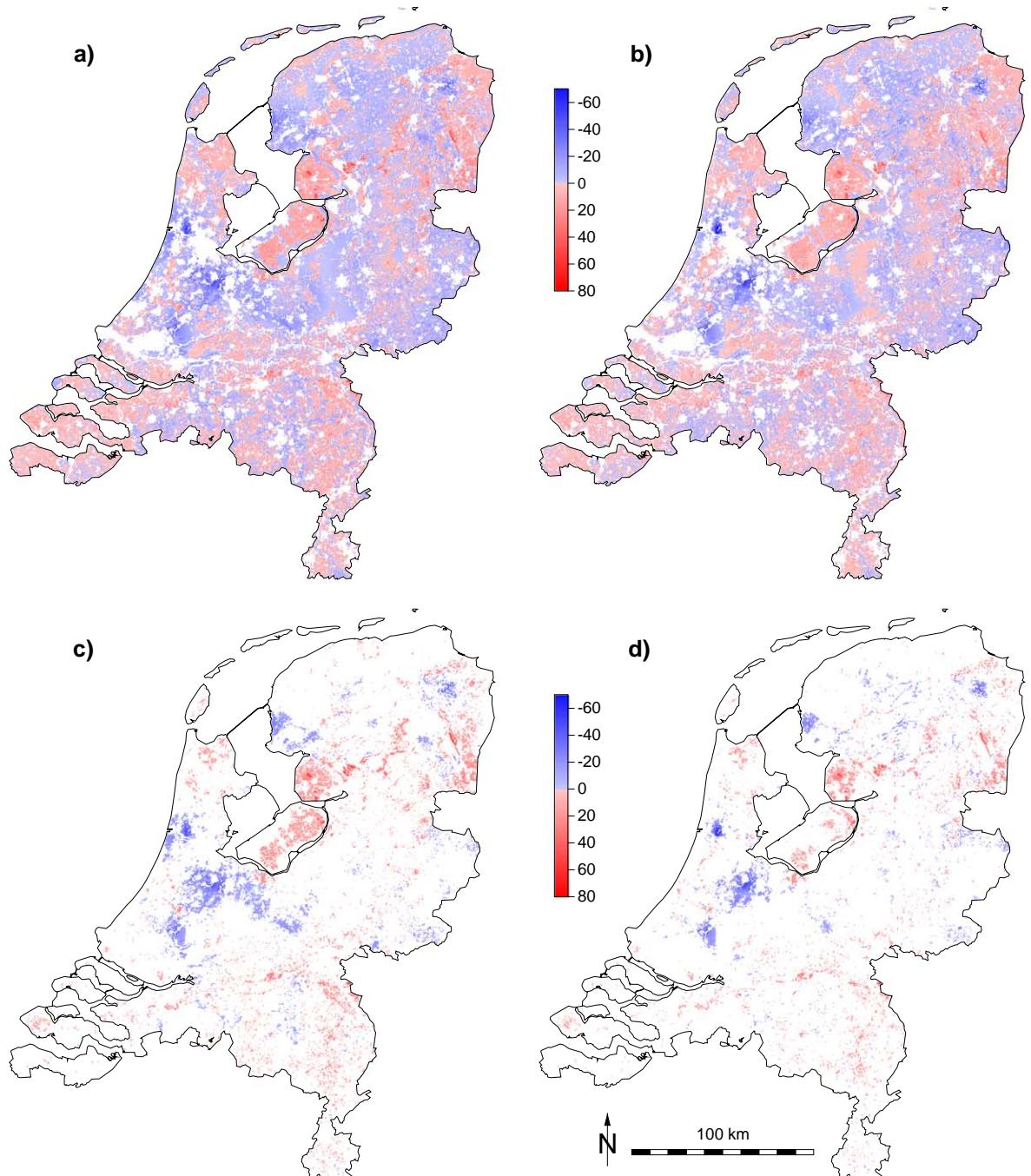


Fig. 7. Residuals for STONE and ordinary block kriging (left) and regression block kriging (right). a) and b) differences for P [mmol/kg soil], c) and d) significant differences for P [mmol/kg soil].

and south). These areas are often characterised by relatively small agricultural fields where landuse can vary substantially over small distances, even within  $500 \times 500$  m grid cells. This local variation is, to some extent, included in the spatial schematisation of STONE but in a different way to the geostatistical predictions. The geostatistical predictions are still relatively smooth in these areas (see Fig. 6a and b).

Fig. 7(c) and (d) show, that for most areas, the prediction error for STONE remains under 40 mmol/kg. Fig. 9(a) and (b) show that the range of absolute differences lies within 10–40 mmol phosphorous per kg soil. STONE systematically overpredicts (red) in the higher range of phosphorous concentrations and systematically underpredicts (blue) in the lower range of phosphorous concentrations. The medium range of STONE predictions are rather equally over- and underpredicted. Some of the under- and overpredictions may be explained by the different degrees of smoothing in the STONE model and kriging, shown by the different variances in Table 3. However, this is not the only explanation for the differences between STONE predictions and kriged

observations. For the OBK comparison significant differences occupy 14.7% of the total area. For the RBK comparison it is 9.8%. This is substantially larger than the 5% probability attributed to the kriging uncertainty.

Fig. 8(a) and (b) show the significant normalised differences. We can see large areas with small to large significant normalised differences and some smaller areas with extremely large significant normalised differences, up to 2000%. A more detailed relationship between STONE predictions and significant differences on the one hand and significant normalised errors on the other hand can be seen in Fig. 9. In the top row the significant differences are plotted against the STONE prediction for the OBK comparison (a) and the RBK comparison (b). Interesting features are the gap around zero of the absolute differences and the triangular shape of the point cloud in the plots. The bottom row shows the corresponding significant normalised differences for the OBK comparison (c) and RBK comparison (d) against the STONE prediction, as computed from Eq. (6). Again, there is a gap around zero. The point clouds show a non-linear shape.

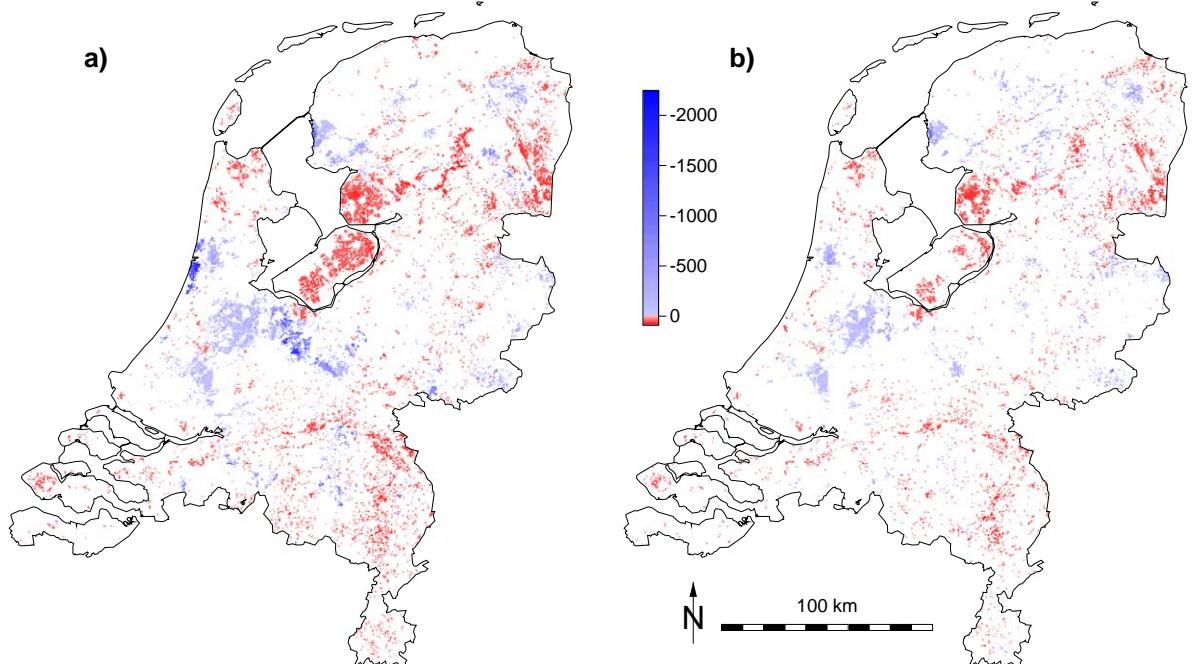


Fig. 8. Significant normalised residuals for STONE and ordinary block kriging (a) and regression block kriging (b) for P [%].

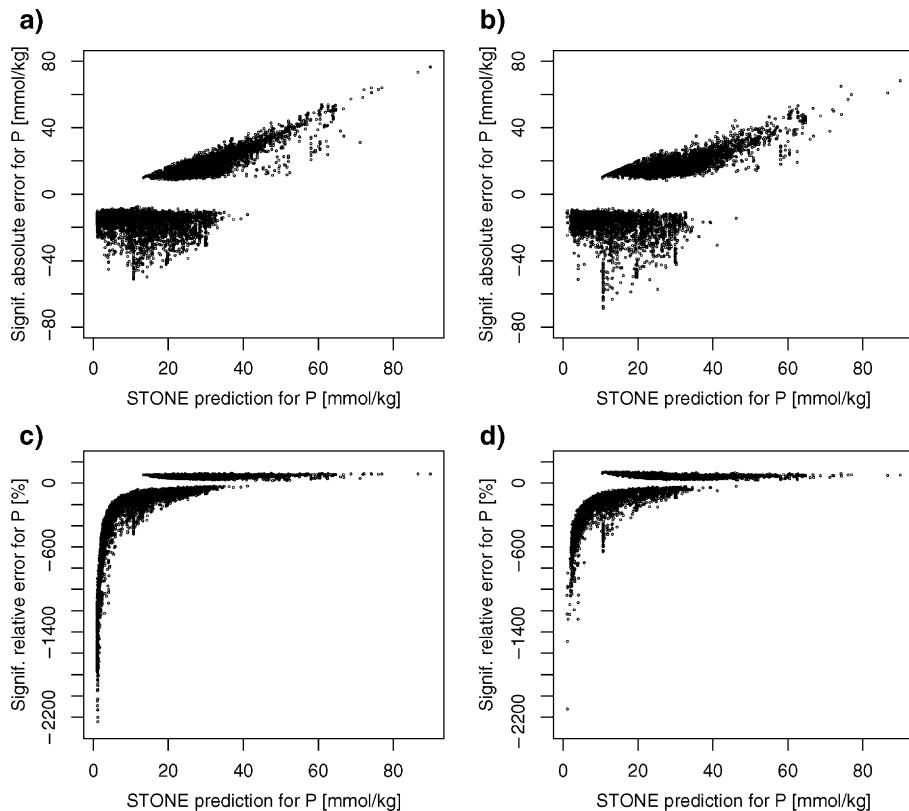


Fig. 9. Scatter plots of STONE predictions against significant differences and against significant normalised differences for ordinary block kriging (a, c) and regression block kriging (b, d).

The normalised differences presented in Fig. 8(a) and (b) show large differences between the areas of overpredictions (red) and underpredictions (blue). Normalised differences are much lower for overpredictions, with maximum values of only 80%, whereas underpredictions approach 2000% in some areas. Here, the residuals are divided by small values from the STONE model, which leads to an amplification of the differences.

Systematic over- and underprediction of up to 40 mmol/kg soil or even larger in some cases could have serious impacts on policy decision. Thus, the errors in predicted P-mod is far from negligible and closer analysis of what the main sources of errors are and how these may be reduced is certainly worthwhile.

It is important to note that by taking +2 and -2 as thresholds for the standardised differences, a wrong conclusion about errors in the STONE model has been

made in about 5% (assuming a normal distribution) of the areas where significant differences are observed. Nonetheless, this study has shown that STONE does over- and underpredict the phosphorous concentration of the top soil for parts of the Netherlands.

#### 4. Conclusions

In this study, the accuracy of simulated mineral phosphorous contents was evaluated by comparing simulation results with 1387 observations from a nation-wide inventory, while accounting for a change in spatial support. The change in spatial support was achieved by interpolating the point observations to the target support using ordinary and regression block kriging. Important advantages of the approach are that the comparison is made at the target support (i.e. the scale of 500 × 500 m grid cells) and that the

interpolation uncertainties are explicitly taken into account. This study has shown that, despite the large number of observations, the interpolation uncertainty was still considerable. One option to reduce interpolation uncertainty could be to add explanatory variables with a higher spatial resolution. Another option could be to use more observations. However, both options are problematic, as there is currently no further detailed information available and collecting additional measurements is time consuming and expensive.

A third option to reduce interpolation uncertainty could be to log-transform the observations prior to the interpolation. But this raises the problem of how to back-transform the interpolated values. One means of overcoming this problem would be to use geostatistical simulation instead of kriging. The point simulated values could then be back-transformed prior to the spatial aggregation. This procedure, however, would need to be repeated a large number of times and the results averaged. Such a Monte Carlo type of simulation approach is computationally demanding.

Despite the fairly large interpolation error, it was possible to demonstrate that the STONE model significantly over- and underpredicts P-mod in parts of the Netherlands. The model appeared to underpredict P-mod in those areas where upward seepage of phosphorous rich water is important and in areas where significant natural background values occur. Overprediction was most probably due to underestimation of transport of manure from farms with a manure surplus. We also observed highly fragmented differences occurring at short distances, due to spatial variability not captured by the model. The high magnitude of underpredictions and overpredictions in some areas and uncertainty in STONE predictions should be taken into account in decision making.

An important problem in the comparison made here is that the regression block kriging errors may be correlated with the STONE errors because the STONE model and regression block kriging make use of the same additional information for soil type and landuse (Heuvelink and Pebesma, 1999). As there is uncertainty associated with the soil and landuse maps, this would mean that errors are introduced in a similar way into both prediction methods, causing correlation between the predictions and underestimation of the

STONE model errors. Although we anticipate that the correlation is small, it may lead to a conservative estimate of the quality of the STONE model predictions. These problems of correlation may be avoided by comparison of the STONE model predictions and observations at the point support, before interpolation and aggregation. The results of the comparison should then be aggregated to the target (block) support, as shown in Fig. 1 by the dashed arrows. This approach is currently being investigated and applied.

The statistical accuracy assessment carried out here yielded significant errors and uncertainties associated with the STONE model predictions at the 500 m block support. For larger supports we would expect the accuracy of STONE to increase. This will be quantified in future research. Another important question is to determine the main cause of error and uncertainty in the STONE model. Answers to these questions can be obtained using uncertainty propagation analyses (Heuvelink, 1998), whereby it is analysed how uncertainties in the various inputs to STONE propagate to the output. Comparison of the model output uncertainty with the propagated input uncertainty will yield insight into whether it is input uncertainty or model uncertainty that is the main source of uncertainty in the STONE predictions.

## Acknowledgements

The work described here was sponsored by the Dutch National Institute of Public Health and the Environment (RIVM). We thank Arthur Beusen (RIVM) for assistance in running STONE and for providing helpful computer scripts. We thank Reind Visschers (ALTERRA) for his support in accessing the LSK data base and James Brown (University of Amsterdam) for valuable comments and improvement of the text.

## References

- Asman, W.A.H., Van Jaarsveld, J.A., 1992. A variable-resolution transport model applied for NH<sub>x</sub> in Europe. Atmospheric Environment. Part A, General Topics 26 (3), 445–464.

- Bierkens, M.F.P., Finke, P.A., De Willigen, P., 2000. Upscaling and Downscaling Methods for Environmental Research. Kluwer, Dordrecht, The Netherlands.
- Bishop, T.F.A., McBratney, A.B., 2001. A comparison of prediction methods for the creation of field-extent soil property maps. *Geoderma* 103 (1–2), 149–160.
- Brown, J.D., Heuvelink, G.B.M., in press. Assessing uncertainty propagation through physically based models of soil waterflow and solute transport. In: Anderson, M.G. (Ed.), Encyclopedia of Hydrological Sciences. Wiley and Sons, Chichester, UK.
- Brus, D.J., de Gruijter, J.J., 1997. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). *Geoderma* 80 (1–2), 1–44.
- Chiles, J.-P., Delfiner, P., 1999. Geostatistics. Modeling spatial uncertainty. Wiley series in probability and statistics. Applied Probability and Statistics. Wiley & Sons, New York.
- Christensen, R., 1990. Linear Models for Multivariate, Time Series and Spatial Data. Springer, New York.
- Cressie, N.A.C., 1993. Statistics for spatial data, revised edition. Wiley series in probability and statistics. Applied Probability and Statistics. Wiley & Sons, New York.
- De Lange, W.J., 1999. A Cauchy boundary condition for the lumped interaction between an arbitrary number of surface waters and a regional aquifer. *Journal of Hydrology* 226 (3–4), 250–261.
- Finke, P.A., de Gruijter, J.J., Visschers, R. 2002. Status 2001: national sampling of map units and applications. Stratified sampling and characterisation of Dutch soils. Tech. Rep., 389 ALTERRA (In Dutch).
- Goovaerts, P., 1997. Geostatistics for natural resources evaluation. Applied Geostatistics Series. Oxford University Press, New York.
- Hengl, T., Heuvelink, G.B.M., Stein, A., 2004. A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma* 120 (1–2), 75–93.
- Heuvelink, G.B.M., 1998. Uncertainty analysis in environmental modelling under a change of spatial scale. *Nutrient Cycling in Agroecosystems* 50 (1–3), 255–264.
- Heuvelink, G.B.M., Pebesma, E.J., 1999. Spatial aggregation and soil process modelling. *Geoderma* 89 (1–2), 47–65.
- Jansen, M.J.W., 1998. Prediction error through modelling concepts and uncertainty from basic data. *Nutrient Cycling in Agroecosystems* 50, 247–253.
- Knotters, M., Brus, D.J., Voshaar, J.H.O., 1995. A comparison of kriging, co-kriging and kriging combined with regression for spatial interpolation of horizon depth with censored observations. *Geoderma* 67 (3–4), 227–246.
- Koopmans, G.F., van der Zeeuw, M.E., Roemkens, P.F.A.M., Chardon, W.J., Oenema, O., 2001. Identification and characterisation of phosphorus-rich sandy soils. *Netherlands Journal of Agricultural Science* 49, 369–384.
- Kroes, J.G., Wesseling, J.C., Van Dam, J.C., 2000. Integrated modelling of the soil–water–atmosphere–plant system using the model SWAP 2.0 an overview of theory and an application. *Hydrological Processes* 14 (11–12), 1993–2002.
- Kroon, T., Finke, P.A., Peereboom, I., Beusen, A.H.W., 2001. Redesign STONE. The new schematization for STONE: the spatial partitioning of land based on hydrological and soil chemical characteristics. Tech. Rep. 2001.017, RIZA, Lelystad, The Netherlands (In Dutch).
- Kros, J., Mol-Dijkstra, J.P., Pebesma, E.J., 2002. Assessment of the prediction error in a large-scale application of a dynamic soil acidification model. *Stochastic and Environmental Research and Risk Assessment* 16, 279–306.
- Loague, K., Corwin, D.L., 1996. Uncertainty in regional-scale assessments of non-point source pollutants. *Soil Science Society of America Journal* 48, 131–152 (SSSA-Special publication).
- Loague, K., Green, R.E., 1991. Statistical and graphical methods for evaluating solute transport models: overview and application. *Journal of Contaminant Hydrology* 7 (1–2), 51–73.
- Overbeek, G.B.J., Tiktak, A., Beusen, A.H.W., van Puijenbroek, P.J.T.M., 2001. Partial validation of the Dutch model for emission and transport of nutrients (STONE). *The Scientific World* 1 (12-S2), 194–199.
- Pebesma, E.J., 2001. Gstat User's Manual, 2nd edition. Utrecht University. (<http://www.gstat.org/gstat.pdf>).
- Pebesma, E.J., Wesseling, C.G., 1998. Gstat, a program for geostatistical modelling, prediction and simulation. *Computers and Geosciences* 24 (1), 17–31.
- R Development Core Team, 2003. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN: 3-900051-00-3 (<http://www.R-project.org>).
- Schoumans, O.F., Groenendijk, P., 2000. Modeling soil phosphorus levels and phosphorus leaching from agricultural land in the Netherlands. *Journal of Environmental Quality* 29 (1), 111–116.
- Shi, W., Fisher, P.F., Goodchild, M.F., 2002. Spatial Data Quality. Taylor and Francis, London.
- Tiktak, A., Beusen, A.H.W., Boumans, L.J.M., Groenendijk, P., de Haan, B.J., Portielje, R., Schotten, C.G.J., Wolf, J., 2003. Verification of STONE version 2.0. Report 718201007. RIVM, Bilthoven, The Netherlands (In Dutch).
- Van Dam, J.C., 2000. Field-scale water flow and solute transport. SWAP Model Concepts, Parameter Estimation and Case Studies. PhD thesis, Wageningen University, Wageningen, The Netherlands.
- Wackernagel, H., 2003. Multivariate Geostatistics, 3rd edition. Springer–Verlag, Berlin.
- Wolf, J., Beusen, A.H.W., Groenendijk, P., Kroon, T., Rötter, R., van Zeijts, H., 2003. The integrated modeling system STONE for calculating nutrient emissions from agriculture in the Netherlands. *Environmental Modelling and Software* 18 (7), 597–617.