Estimation of the impact of short-term fluctuations in inputs on temporally aggregated outputs of process-oriented models

Å. Forsman, C. Andersson, A. Grimvall and M. Hoffmann

ABSTRACT

Process-oriented models driven by highly resolved meteorological inputs and comprising a short internal time step are sometimes used to predict substance fluxes in air, soil and water over fairly long periods of time. To ascertain whether regression-based input–output analyses in such cases can provide adequate parametric models of the impact of daily and monthly fluctuations in inputs on annual outputs, we studied the *SOIL/SOILN* model of vertical transport of heat, water and nitrogen through arable soils. Annual leaching of nitrate from the root zone was regarded as the response variable, and regressors were selected from among the set of all linear combinations of daily or monthly values of five different meteorological inputs. We found that, although several of the underlying processes described by the *SOIL/SOILN* model are non-linear, both ordinary and partial least squares regression (OLS and PLS) identified the subsets of input variables with the strongest influence on the model output, and the dominating time lags between model inputs and outputs. Furthermore, highly resolved explanatory variables were a prerequisite for good performance of linear predictors of temporally aggregated outputs and, to discern the full dynamic behaviour of the model, it was necessary to analyse the response to artificially generated daily meteorological data representing a very large number of different weather conditions. PLS had one advantage over OLS: a smooth pattern in the regression coefficients facilitated physical interpretation of the derived impulse–response weights.

Key words | deterministic models, nitrate leaching, partial least squares, process-oriented models, regression analysis, temporal aggregation.

Å. Forsman (corresponding author)
Department of Mathematics,
Linköping University,
SE-58183 Linköping,
Sweden
E-mail: *asa.forsman@vti.se*

C. Andersson
Department of Mathematics,
Kaiserslautern University,
D-67653 Kaiserslautern,
Germany

A. Grimvall
Department of Mathematics,
Linköping University,
SE-58183 Linköping,
Sweden

M. Hoffmann
Department of Soil Sciences,
PO Box 7072,
Swedish University of Agricultural Sciences,
SE-75007 Uppsala,
Sweden

INTRODUCTION

In process-oriented, deterministic models of environmental systems, the output is uniquely determined by the initial state of the studied system, the inputs and a set of model parameters. Nevertheless, it can be difficult to comprehend how fluctuations in the inputs influence the outputs. For example, it is often practically impossible to trace the impact of natural fluctuations in weather through each of the different processes and compartments of a studied system. Hence, there is a strong need for procedures that can extract simplicity out of complexity and thereby render models of environmental systems more transparent (Young *et al.* 1996). Model simplifications have been found to be particularly valuable when models developed for small spatial units are to be extrapolated to large areas (e.g. Bouzaher *et al.* 1993, de Vries *et al.* 1998), and for incorporating the knowledge gained from process-oriented modelling into decision support tools (Quinn *et al.* 1999, Forsman *et al.* 2002a).

The present study was focused on elucidating the influence of highly resolved inputs on the total outputs over periods that are much longer than the internal time step of the model under consideration. It is easy to show that such aggregation of outputs can enable considerable simplification of many process-oriented models. First, some inputs that mainly influence the short-term dynamics of the outputs may be omitted. Secondly, temporally aggregated outputs can be almost linear functions of the inputs, even though several of the processes included in the model are highly non-linear (Forsman & Grimvall 2002). Impulse–response weights in linear models are also highly interesting because they are easy to comprehend, whereas parameters of non-linear models, such as artificial neural networks (ANNs), can rarely be given a physical interpretation (Dawson & Wilby 2001). Hence, we examined the use of regression analysis of large sets of model inputs and outputs to estimate impulse–response weights that describe the impact of highly resolved inputs on temporally aggregated outputs.

Inasmuch as the total model output for a certain period, for example one year, can be influenced by a very large number of daily or monthly inputs for the current and previous years, we need statistical procedures that can clarify the response to a substantial number of strongly correlated explanatory variables. Over the past decades a variety of regression methods, such as principal components regression, ridge regression and partial least squares regression (PLS), have come to be widely used to resolve such issues, and there is now a unified theory regarding these techniques. In particular, it has been demonstrated that there is a continuous spectrum of regression methods that provide a link from principal components regression, over PLS, to ordinary least squares regression (OLS) (Stone & Brooks 1990). The relationship between PLS and ridge regression has also been clarified (Sundberg 1993, Björkström & Sundberg 1999).

In a previous study (Forsman *et al.* 1998), we used PLS to elucidate the dynamic behaviour of the soil nitrogen model *SOIL/SOILN* (Johnsson *et al.* 1987, Jansson & Halldin 1979), which is a process-oriented model of the vertical transport of heat, water and nitrogen through arable soils. More specifically, we showed that statistical analysis of the response of the model to artificially generated meteorological inputs could explain the connection between annual totals of nitrate leaching and monthly mean values of air temperature, precipitation and other meteorological variables. The present study was devoted to a more thorough analysis of the feasibility of using linear statistical approximations of basically non-linear models to reveal possible effects of short-term fluctuations in the inputs on temporally aggregated model outputs. The focus was on using artificially generated inputs and outputs of the *SOIL/SOILN* model to investigate the following:

(i) the amount of data needed to discern statistical relationships between the annual leaching of nitrate and daily or monthly meteorological inputs;
(ii) the feasibility of handling regression models with up to a thousand explanatory variables, corresponding to different meteorological inputs at different times;
(iii) the goodness-of-fit of linear models based on either daily or monthly meteorological inputs;
(iv) the possible advantages of PLS over OLS for analysis of input–output data.

MODELS AND DATA

The *SOIL/SOILN* model

The *SOIL/SOILN* model comprises a soil water and heat module (Jansson & Halldin 1979) and a nitrogen module (Johnsson *et al.* 1987) coupled in series. The water and heat module uses daily meteorological data (air temperature, cloudiness, precipitation, vapour pressure and wind speed) as inputs to predict soil water and heat conditions at any level in a soil profile; the main equations are derived from Fourier's and Darcy's laws. The nitrogen module includes the major processes that determine inputs, transformations and outputs of nitrogen in arable soils (Figure 1). Nitrogen inputs can be in the form of commercial fertiliser or manure added to the topsoil or atmospheric deposition; harvesting, leaching, and denitrification constitute the outputs. Litter, faeces and humus represent different fractions of organic nitrogen. Moreover, organic carbon pools are included for litter and faeces in order to regulate nitrogen mineralisation.

The general structure of the *SOIL/SOILN* model enables simulation of nitrate leaching from a great variety of cropping systems. The model parameters in our study were selected to represent cultivation of barley on a sandy soil in southern Sweden. Commercial fertiliser was

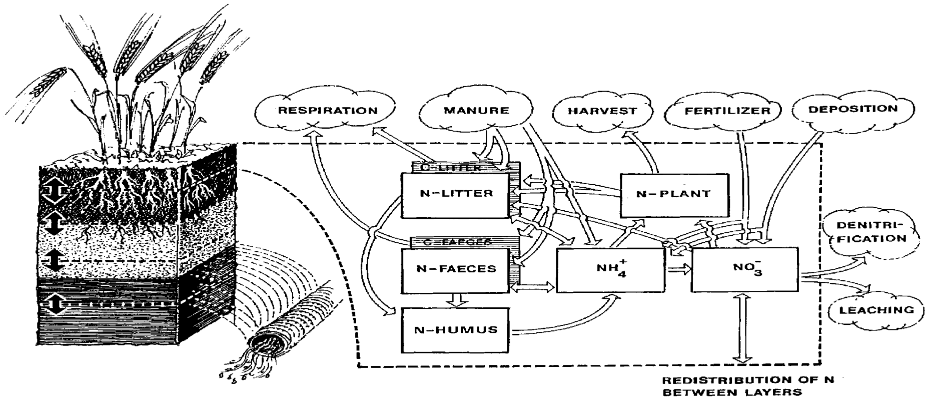

Figure 1 | State variables (boxes) and flows (arrows) included in the *SOILN* model. Boxes and arrows enclosed by the dashed line represent the top layer of the soil. The lower layers have the same structure but have no direct input in the form of fertilisers or atmospheric deposition. Source: Johnsson *et al.* (1987).

applied to the topsoil once a year in the middle of April, at the same time as the barley was sown. The crops were harvested in the middle of August.

Observed and synthetic meteorological data

Daily meteorological records from a station located close to the city of Lund in southern Sweden were obtained from the Swedish Meteorological and Hydrological Institute. The observation period ranged from 1961 to 1994, and data were compiled for the following variables:

T = air temperature (°C)
C = relative cloudiness
P = precipitation (mm d^{-1})
V = vapour pressure (Pa)
W = wind speed (m s^{-1})

The observed time series of data can be considered to be a realisation of a multivariate stochastic process. To be able to generate other realisations of the same process, it is necessary to identify the underlying multivariate probability distributions. We used vector autoregressive models to generate daily data (Forsman *et al.* 2002b) and multivariate regression models to generate monthly data. Seasonal fluctuations were taken into account by generating data separately for each month of the year. Non-normality was handled by transforming the original data, fitting a model to the transformed data and finally transforming the generated data back to the original form.

REGRESSION METHODS

Feeding the *SOIL/SOILN* model with *observed* or *synthetic* meteorological data produced values of the selected response variable, that is, annual nitrate leaching. Monthly or daily meteorological data for the current and previous years were selected as explanatory variables in the regression analysis. To enable identification of the variables that had the strongest influence on the model output, prior to the analysis, we standardised the data for each meteorological variable to unit mean standard deviation, where the mean was taken over the monthly values. The results are presented as impulse–response weights (regression coefficients) for the standardised explanatory variables. All data were analysed by both PLS and OLS.

PLS is an indirect regression technique in which the variation of a response variable is linked to a large number of explanatory variables through a small or moderate number of factors that are defined as normed linear combinations of the explanatory variables. The first version of PLS was described as a numerical algorithm (Wold 1975). The theoretical aspects of this method have now been

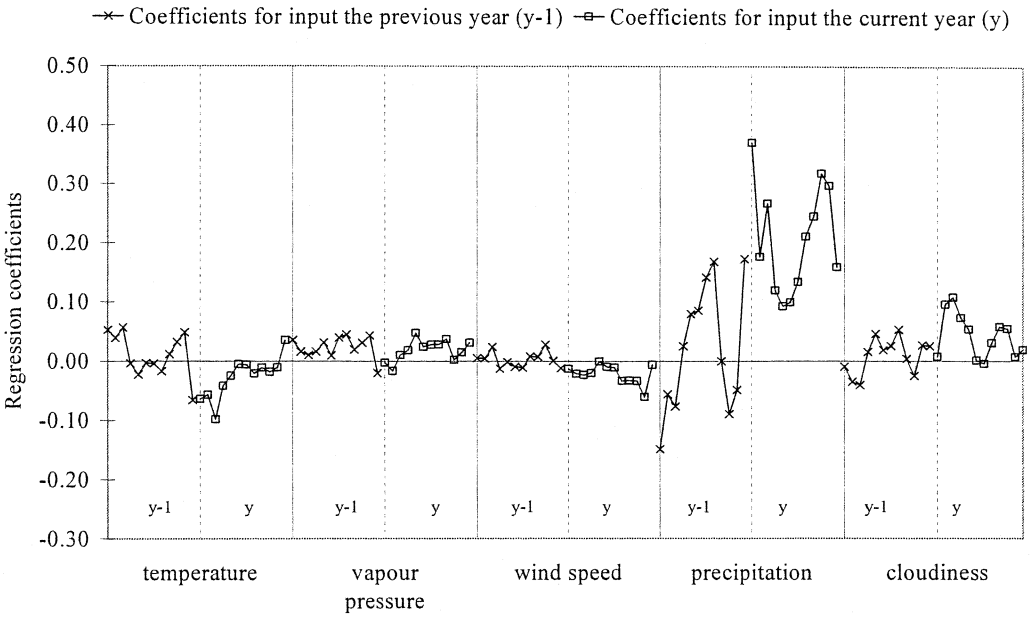

Figure 2 | PLS analysis of annual nitrate leaching in response to monthly fluctuations in *synthetic* meteorological variables. The curve illustrates regression coefficients obtained by applying a three-factor PLS model to analyse data representing a period of 400×30 years. The current and previous year are respectively denoted *y* and *y*−1 and the coefficients in each group represent the twelve months of the year.

thoroughly investigated (Frank 1987, Helland 1988, 1990, Höskuldsson 1988, Garthwaite 1994) and PLS has become a standard tool in chemometrics and multivariate calibration (Martens & Naes 1989, Brown 1993).

The first factor in a PLS model is selected to maximise the covariance with the response variable and is subsequently used as a regressor in an OLS regression model. The next factor is selected to maximise the covariance with the estimated residuals from the OLS model.

The results of the final PLS model are presented as estimates, $\hat{\boldsymbol{b}}_{\mathrm{PLS}}$, of the regression coefficients, $\boldsymbol{b}$, in the linear regression model

$$\boldsymbol{y} = \bar{\boldsymbol{y}} \quad + (\boldsymbol{X} - \bar{\boldsymbol{X}} \quad)\boldsymbol{b} + \boldsymbol{e}$$

where $\boldsymbol{y}$ is a vector containing the response values, $\boldsymbol{X}$ is a matrix of the explanatory variables and $\boldsymbol{e}$ is a vector of uncorrelated errors with equal variance.

The number of factors to be used in the final PLS model can be determined by employing cross-validation to study the ability of different models to predict the response variable. In the present study, the observations were divided into two data sets: one was used to fit the model and the other to validate the model by comparing predicted and true values. The prediction error sum of squares (PRESS) was calculated for different numbers of factors; the number of factors used in the model was no longer increased when the addition of a new factor resulted in only a small decrease in the PRESS value.

RESULTS

Analysis of the response to monthly fluctuations in meteorological data

The results of PLS analysis of the response to monthly fluctuations in synthetic meteorological data are shown in Figure 2. Each meteorological variable generates 24 explanatory variables in the regression model. Considering cloudiness as an example, the markers in Figure 2 (from left to right) represent January of the previous year to December of the current year. The pattern

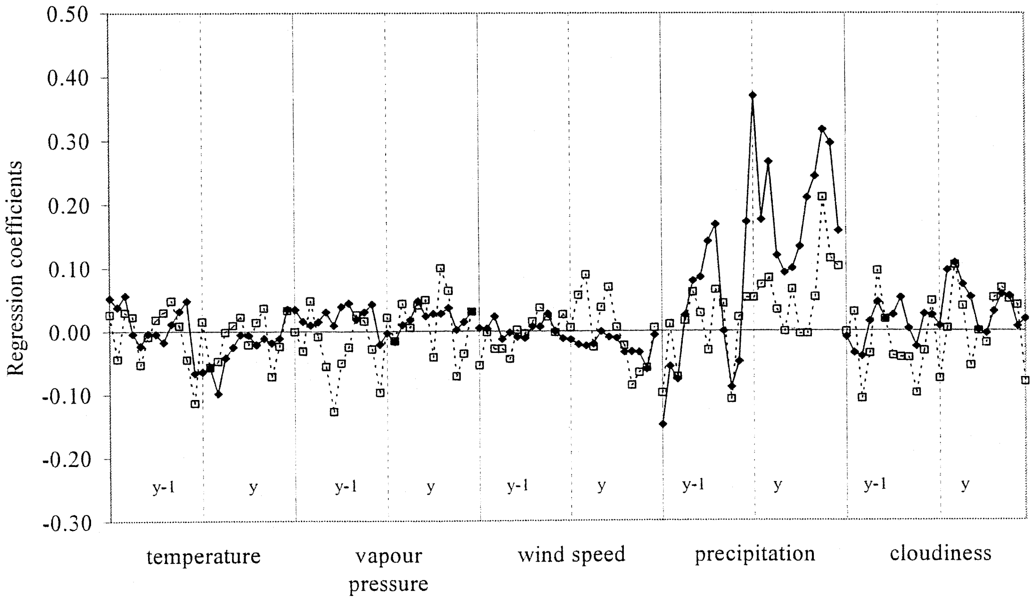

Figure 3 | PLS analysis of annual nitrate leaching in response to monthly fluctuations in *synthetic* (solid line) and *observed* (dashed line) meteorological variables. The curve illustrates regression coefficients obtained by applying a three-factor PLS model.

of the regression coefficients illustrates that the variables that had the greatest impact on the annual leaching of nitrate were precipitation, followed by cloudiness and air temperature. Moreover, there were rather simple mechanistic explanations for the most pronounced patterns in the estimated regression coefficients. The positive regression coefficients for monthly precipitation during the current year reflect the obvious fact that a particularly heavy rain or snowfall will result in elevated leaching of nitrate the same year, and the seasonal pattern in the same coefficients indicates that the percentage of precipitation that generates runoff is lower in summer than in the other parts of the year. Closer examination of the regression coefficients revealed that there were also plausible explanations for minor details. For example, the comparatively low value of the coefficient for precipitation in February of the current year can be explained by the accumulation of snow, whereas the negative coefficients for precipitation in October and November of the previous year suggest that, after a heavy autumn rain, there is less nitrate in the soil that can be washed out by subsequent rainfalls.

In an attempt to estimate the amount of data needed to identify the most influential input variables, we compared two PLS models, one based on 30 years of *observed* meteorological data and the other on 400 × 30 years of *synthetic* meteorological data. Figure 3 shows that the large set of *synthetic* input data produced a pronounced pattern in the regression coefficients, whereas the *observed* meteorological data resulted in more irregular variation in the coefficients. Closer examination of the regression coefficients obtained for different subsets of *synthetic* meteorological data (Figure 4) showed that 30 years of such data do not provide estimates stable enough to reveal the most influential input variables or the major time lags between inputs and outputs.

Comparison of a three-factor PLS model and an OLS model is illustrated in Figure 5. The pattern of the regression coefficients is more irregular with OLS than with PLS, especially regarding the variables of temperature and vapour pressure. To determine whether this difference between the OLS and PLS coefficients was due to uncertainty in the parameter estimates, we divided the total data set into four subsets of equal size and then estimated the coefficients for each subset. The results, presented in Figures 6 and 7, indicate slightly larger variability in the OLS estimates than in the PLS estimates, but

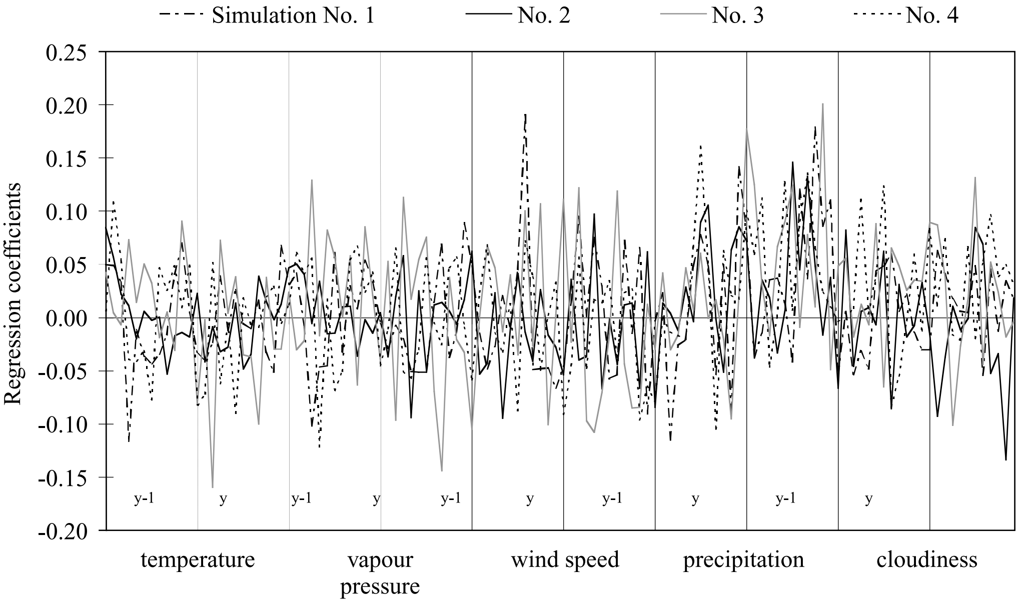

Figure 4 | PLS analysis of annual nitrate leaching in response to monthly fluctuations in *synthetic* meteorological variables. Each of the four curves illustrates the regression coefficients obtained by applying a three-factor PLS model to analyse data representing a period of 30 years.

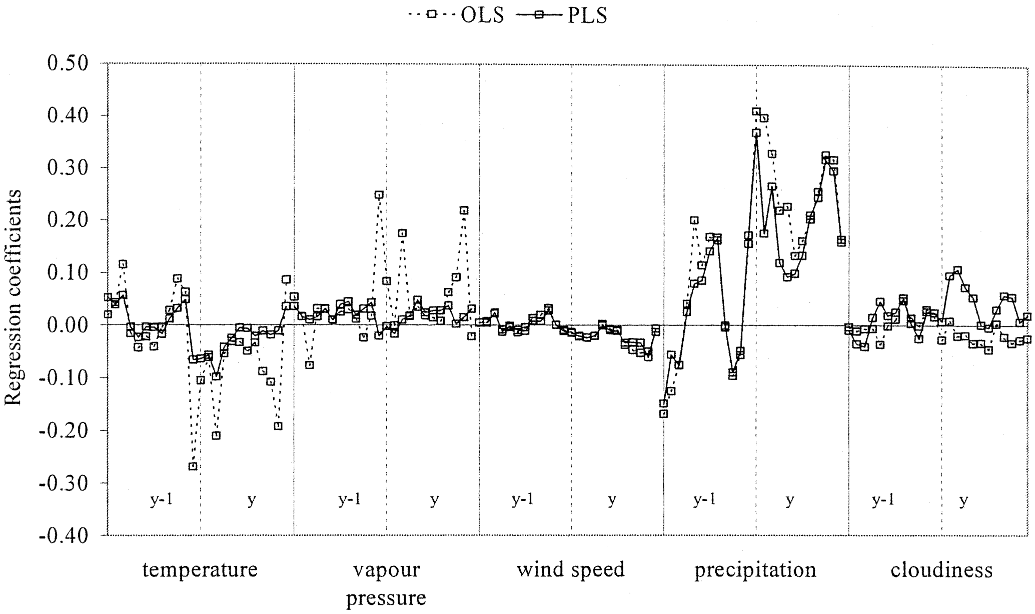

Figure 5 | OLS and PLS analysis of annual nitrate leaching in response to monthly fluctuations in *synthetic* meteorological variables. The regression coefficients shown were obtained by a three-factor PLS analysis (solid line) and OLS analysis (dashed line).

the same overall pattern is seen for each regression model in all four estimates. Notably, regression coefficients for strongly correlated explanatory variables, such as contemporaneous temperature and vapour pressure, were approximately equal in the PLS analysis but were in some cases markedly different in the OLS analysis.

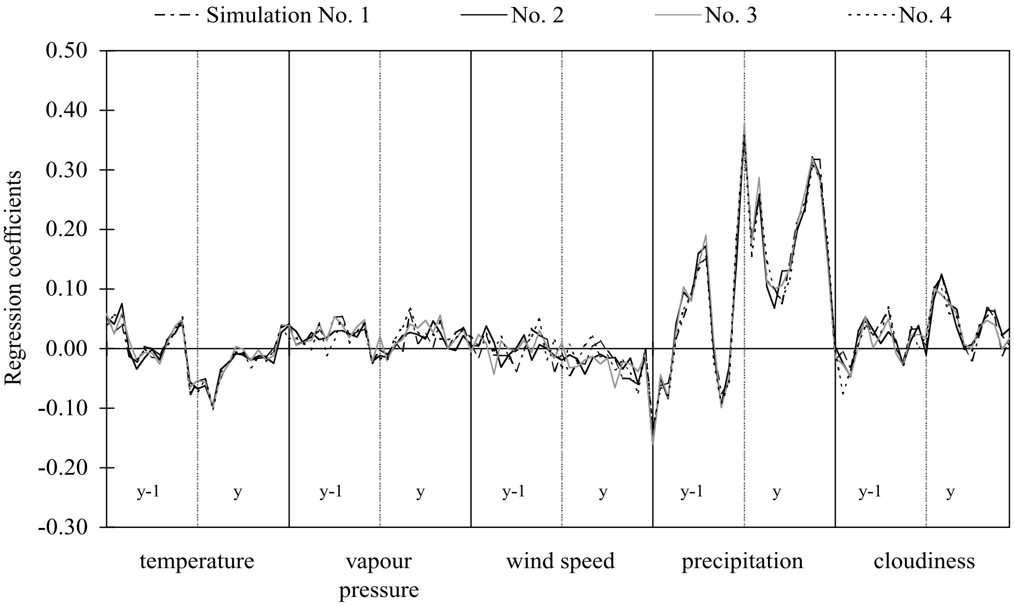

Figure 6 | PLS analysis of annual nitrate leaching in response to monthly fluctuations in *synthetic* meteorological variables. Each of the four curves illustrates the regression coefficients obtained by applying a three-factor PLS model to data representing a period of 100×30 years.

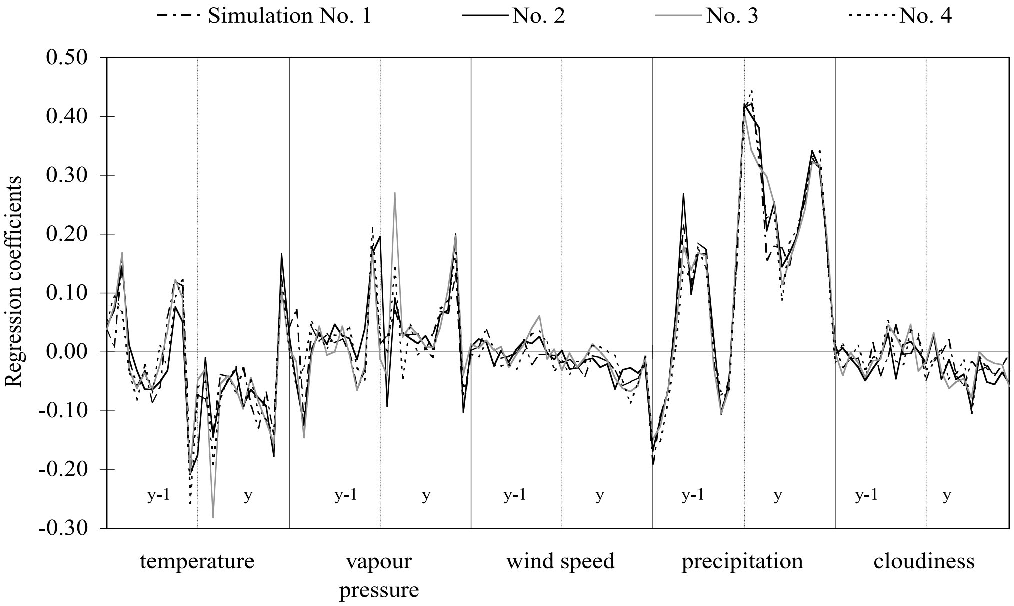

Figure 7 | OLS analysis of annual nitrate leaching in response to monthly fluctuations in *synthetic* meteorological variables. Each of the four curves illustrates the regression coefficients obtained by applying an OLS model to data representing a period of 100×30 years.

Figure 8 illustrates predicted versus simulated nitrogen leaching for two OLS models. In the first model (a), annual nitrate leaching is used as the response variable, and monthly meteorological data for two years are selected as explanatory variables. The second model (b) describes nitrogen leaching aggregated to two-year means

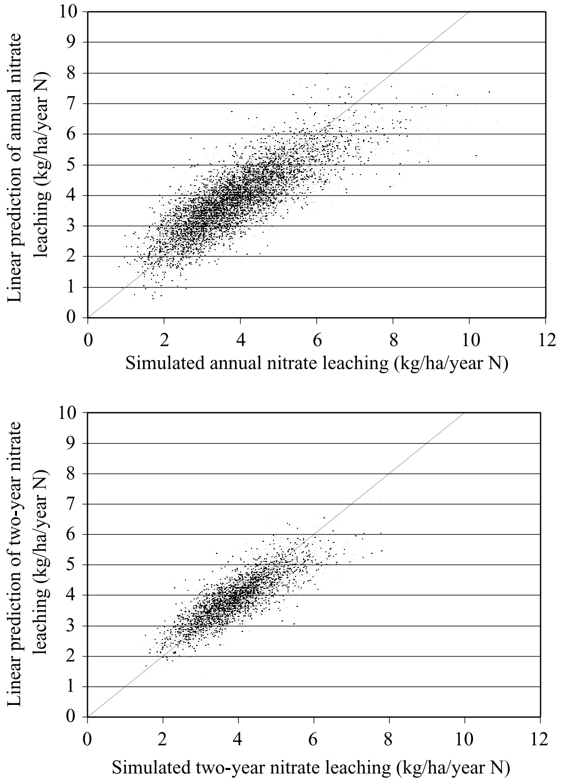

Figure 8 | Regression analyses of temporally aggregated nitrogen leaching versus monthly meteorological data for the corresponding time period and one previous year. The diagrams show predicted versus simulated nitrogen leaching when the response is aggregated to one-year (a) and two-year (b) values. The analysis in (a) is based on 12,000 observations and 120 explanatory variables, and the analysis in (b) on 5,800 observations and 180 explanatory variables.

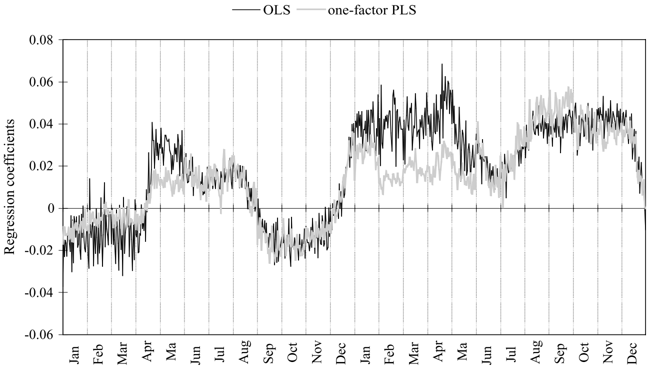

Figure 9 | OLS and PLS analysis of annual nitrate leaching in response to daily fluctuations in *synthetic* precipitation. The regression coefficients shown were obtained by a one-factor PLS analysis (shaded line) and by OLS analysis (solid line).

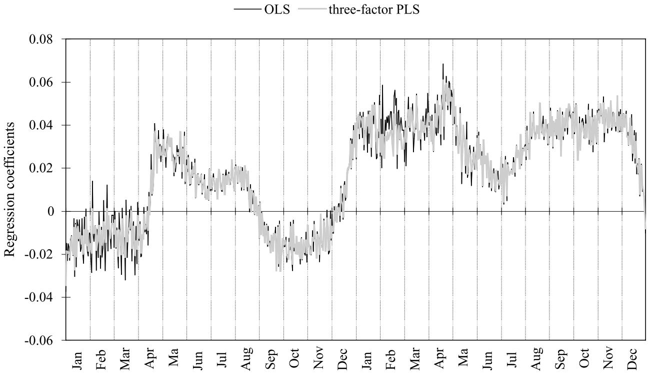

Figure 10 | OLS and PLS analysis of annual nitrate leaching in response to daily fluctuations in *synthetic* precipitation. The regression coefficients shown were obtained by a three-factor PLS analysis (shaded line) and by OLS analysis (solid line).

in response to monthly meteorological data representing a period of three years.

Since the output from the *SOIL/SOILN* model is determined entirely by the input variables, there are three different reasons for the differences between estimated and observed nitrogen leaching: (i) the temporal resolution of the input data has been reduced from daily values in the model simulations to monthly values in the regression analyses; (ii) non-linear parts of the relationship are not captured by the regression models; (iii) the output from *SOIL/SOILN* for a specific year depends on all previous inputs, and we restricted the time-lagged data in our study to a single previous year.

Figure 8 shows no severe non-linearities in the relation between the estimated and observed nitrogen leaching. There is only slight non-linearity when the response is annual nitrogen leaching (Figure 8a). In particular, large simulated values from the *SOIL/SOILN* model are underestimated by the regression model. When nitrogen leaching is aggregated to two-year means (Figure 8b), the response becomes even more linear and the residuals are smaller. The R^2 values for the models in Figure 8(a, b) are 0.67 and 0.70, respectively.

Analysis of the response to daily fluctuations in meteorological data

The number of explanatory variables increased dramatically when the analysis of the response to fluctuations in

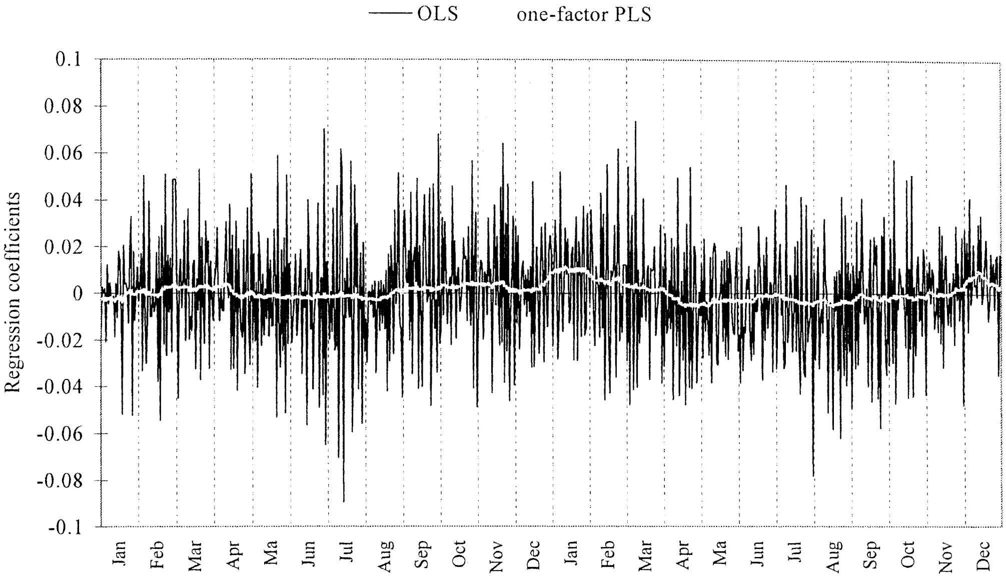

Figure 11 | OLS and PLS analysis of annual nitrate leaching in response to daily fluctuations in *synthetic* temperature. The regression coefficients shown were obtained by a one-factor PLS analysis (shaded line) and by OLS analysis (solid line).

meteorological data was extended from monthly averages to daily data, and so we analysed the meteorological variables separately. Figure 9 shows the nitrate leaching response to daily fluctuations in precipitation for an OLS model and a one-factor PLS model. The PLS coefficients appear to be biased, but the overall pattern is similar for the two methods. With three factors in the PLS analysis, the regression coefficients are almost the same (see Figure 10). In Figure 11, daily temperatures are the explanatory variables and the differences between the two methods are more evident due to the large autocorrelations in temperature, as compared to precipitation. Closer examination of the larger variability in the OLS estimates indicated that this was due to considerable statistical uncertainty; when the regression coefficients were estimated separately for two subsets of the original data, there were distinctly different patterns in the coefficient estimates.

Predicted versus simulated nitrate leaching for two models based on daily inputs is depicted in Figure 12. Both of the illustrated models predict two-year means of nitrate leaching; explanatory variables are three years of daily precipitation for the first model and three years of daily precipitation and temperature for the second model. The R^2 values for the two models are 0.81 and 0.87, respectively. Accordingly, employing daily instead of monthly data clearly improves the models, and this is even more evident considering that the models with daily data do not include all meteorological inputs.

The impact of smoothing day-to-day variation in meteorological data

When using monthly averages of the meteorological data as explanatory variables, the required daily input to the *SOIL/SOILN* model was generated by spreading the monthly precipitation uniformly over the days of the month. We have already seen that such smoothing of the model input resulted in lower R^2 values. Further analysis demonstrated that the cumulative values of nitrate leaching obtained in *SOIL/SOILN* simulations decreased significantly when the day-to-day variation in meteorological inputs was removed (see Figure 13). Hence, smoothing of the inputs can also jeopardise the physical interpretation of the coefficients derived by regressing annual leaching on monthly averages of the meteorological inputs.

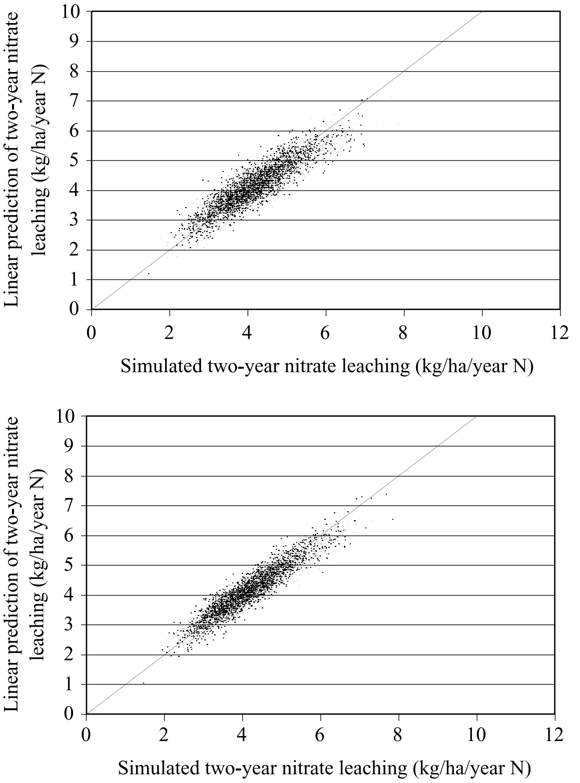

Figure 12 | Regression analyses of temporally aggregated nitrogen leaching versus daily meteorological data for the corresponding time period and one previous year. The diagrams show predicted versus simulated two-year nitrate leaching when the explanatory variables are precipitation (a) or precipitation and temperature (b). The analyses are respectively based on 1,095 and 2,190 explanatory variables and 6,000 observations.

Hydrologists have long known that, for a given amount of precipitation, the water flow in soil will be lower if there is the same amount of precipitation every day rather than larger amounts on a limited number of days. Since nitrate is dissolved and transported in the water, the reduced water flow will also lead to a decrease in nitrate leaching.

DISCUSSION

The present results demonstrate that regression analysis of input–output data can explicitly explain the effects of random weather fluctuations on annual outputs of process-oriented models driven by daily or monthly meteorological data. Provided the available data represented a sufficient number of different weather conditions, the derived impulse–response weights allowed identification of both the most influential input variables and the characteristic time lags between inputs and outputs. Moreover, as indicated in our comments on Figure 2, there were plausible mechanistic interpretations of both major and minor features of the pattern of impulse–response weights.

Superficially, non-linear models such as artificial neural networks (ANNs) seem to provide appealing solutions to the problem of relating outputs to inputs of complex models (Govindaraju & Ramachandra Rao 2000, Dawson & Wilby 2001). However, it has also been recognised that linear methods such as PLS can be competitive (Hadjiiski *et al.* 1999). In the introduction, it was mentioned that physical interpretation of impulse–response weights in a linear model is often possible, whereas non-linear models such as neural networks have a pronounced black-box character. Secondly, we found that temporally aggregated model outputs could be accurately predicted by linear expressions in highly resolved inputs, even though the model under consideration involved several markedly non-linear processes. For example, we obtained an R^2 value of 87% when the total leaching of nitrogen over a period of two years was regressed on daily precipitation and temperature values. Thirdly, there are linear regression techniques that have been designed specifically to handle a substantial number of strongly correlated predictors, whereas, in such cases, ANNs and other procedures involving very large classes of models may lead to overfitting if extensive precautions are not taken during model identification (Bishop 1995). Finally, it should be mentioned that the performance of linear predictors involving highly resolved (daily) inputs was superior to that of linear predictors based on temporally aggregated (monthly) inputs.

The two regression methods we investigated, PLS and OLS, produced regression coefficients that were almost identical for the variable precipitation but differed greatly for the other variables, especially temperature and vapour pressure. Closer inspection of the results presented in Figure 5 revealed that the major differences appeared when two or more explanatory variables were strongly correlated to each other but weakly correlated to the

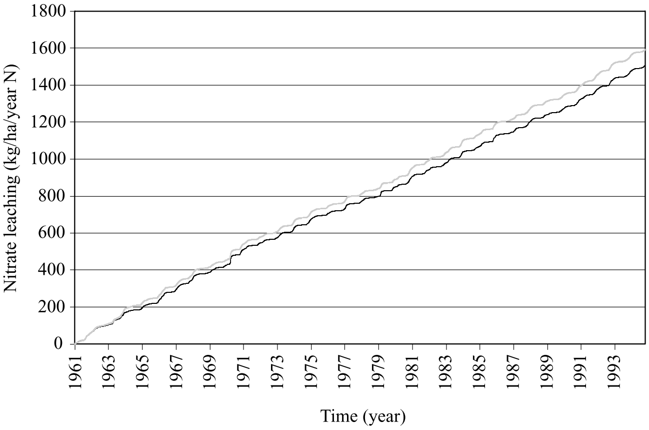

Figure 13 | Cumulated annual nitrate leaching in *SOIL/SOILN* simulations carried out using observed daily meteorological data (dashed line) and monthly averages (solid line).

response variable. For example, the largest negative coefficients for temperature in the OLS analysis are seen for the months with strongly positive coefficients for vapour pressure. The patterns of the PLS coefficients were smoother, because, in that type of regression, strongly correlated explanatory variables normally result in almost identical regression coefficients. Smooth patterns are easier to interpret in terms of underlying mechanisms: thus PLS is superior to OLS, if the main objective of the input–output analysis is to reveal the dynamic properties of a complex mechanistic model. Furthermore, it should be stressed that smoothing the OLS coefficients over time will not produce PLS-like coefficients, since there may be strong correlations both within and between meteorological variables. Principal components analysis was tested during an initial stage of the work but was subsequently abandoned, because a large number of principal components were needed to adequately represent the total variation in the input variables. Ridge regression would most probably be a viable alternative to PLS (Sundberg 1993, Björkström & Sundberg 1999), but a comparison of those two techniques was outside the scope of the present study.

When a parametric model is derived solely for the purpose of prediction, a set of input–output data of moderate size may suffice. The present investigation was also focused on the parameters of the fitted model and, in such cases, much larger data sets are needed. We noted that a readily interpretable pattern in the regression coefficients did not emerge until considerable amounts of input–output data were analysed. When data representing a period of 'only' 30 years were used, the uncertainty of the estimated coefficients was so large that it was difficult to identify even the most influential input variables (Figure 4). This implies that, in practice, artificially generated weather data must be employed to reveal the dynamic properties of models as complex as *SOIL/SOILN*.

Artificial weather data can be generated in many different ways. For example, a large number of new time series can be obtained by resampling of a set of observed data. Our approach was slightly more sophisticated in the sense that it enabled generation of new weather events that were consistent with observed univariate distributions of each of the meteorological variables, and that it had realistic cross- and autocorrelations. However, it should be emphasised that we made no attempt to take into account the uncertainty involved in estimating the multivariate distribution of the different weather variables.

CONCLUSIONS

The impact of short-term fluctuations in inputs on temporally aggregated outputs can be estimated by regression analysis of large sets of model inputs and outputs.

High resolution of model inputs is a prerequisite of obtaining a good linear approximation of the process-oriented model.

Both PLS and OLS permit identification of the meteorological variables that have the strongest impact on model outputs and the characteristic time lags between model inputs and outputs.

If the input variables are strongly correlated, smoother regression coefficient patterns will be produced by PLS than by OLS. Thus PLS will facilitate the search for mechanistic explanations for derived patterns of regression coefficients.

REFERENCES

Bishop, C. M. 1995 *Neural Networks For Pattern Recognition*. Clarendon Press, Oxford.

Björkström, A. & Sundberg, R. 1999 A generalized view on continuum regression. *Scand. J. Stat.* **26**, 17–30.

Brown, J. P. 1993 *Measurement, Regression, and Calibration*. Clarendon Press, Oxford.

Bouzaher, A., Lakshiminarayan, P. G., Cabe, R., Carriquiry, A., Gassman, P. W. & Shogren, J. F. 1993 Metamodels and nonpoint pollution policy in agriculture. *Wat. Res. Res.* **29** (6), 1579–1587.

Dawson, C. W. & Wilby, R. L. 2001 Hydrological modelling using artificial neural networks. *Prog. Phys. Geog.* **25** (1), 80–108.

de Vries, W., Kros, J., van der Salm, C., Groenenberg, J. E. & Reinds, G. J. 1998 The use of upscaling procedures in the application of soil acidification models at different spatial scales. *Nutr. Cycl. Agroecosys.* **50**, 223–236.

Forsman, Å., Andersson, C., Grimvall, A. & Hoffmann, M. 1998 Partial least squares (PLS) regression to extract simple statistical relationships from complex environmental models. In: *Second International Symposium on Sensitivity Analysis of Model Output (SAMO 98), Venice, Italy* (ed. Chan, K., Tarantola, S. & Campolongo, F.). European Commission. pp. 115–118.

Forsman, Å., Grimvall, A., Scholtes, J. & Wittgren, H. B. 2002*a* Generic structures of decision support systems for evaluation of policy measures to reduce catchment-scale nitrogen fluxes. In: *Identification of simple structures in complex substance transport models.* PhD Thesis, Linköping Studies in Statistics No. 1, Linköping, Sweden.

Forsman, Å., Björklund, C. & Grimvall, A. 2002*b* Simulation of multivariate time series of meteorological data. Research report LiU-MAT-R-2002-02, Linköping University, Linköping, Sweden.

Forsman, Å. & Grimvall, A. 2002 Linearisation of highly resolved substance transport models by temporal aggregation of model outputs. In: *Proceedings of the International Conference on Integrated Assessment and Decision Support (iEMSs 2002), Lugano, Switzerland, 24-27 June, 2002* (ed. Rizzoli, A. E. & Jakeman, A. J.). iEMSs, 375–380.

Frank, I. E. 1987 Intermediate least squares regression method. *Chemometr. Intell. Lab.* **1**, 233–242.

Garthwaite, P. H. 1994 An interpretation of partial least squares. *J. Am. Statist. Assoc.* **89**, 122–127.

Govindaraju, R. S. & Ramachandra Rao, A. (eds) 2000 *Artificial Neural Networks In Hydrology.* Kluwer, Dordrecht.

Hadjiiski, L., Geladi, P. & Hopke, P. 1999 A comparison of modeling nonlinear systems with artificial neural networks and partial least squres. *Chemometr. Intell. Lab.* **49**, 91–103.

Helland, I. S. 1988 On the structure of partial least squares regression. *Commun. Stat. B Simul.* **17**, 581–607.

Helland, I. S. 1990 Partial least squares regression and statistical models. *Scand. J. Stat.* **17**, 97–114.

Höskuldsson, A. 1988 PLS regression methods. *J. Chemometr.* **2**, 211–228.

Jansson, P. E. & Halldin, S. 1979 Model for annual water and energy flow in a layered soil. In: *Comparison of Forest Water and Energy Exchange Models* (ed. Halldin, S.). International Society for Ecological Modelling, Copenhagen, pp. 145–163.

Johnsson, H., Bergström, L., Jansson, P. E. & Paustian, K. 1987 Simulated nitrogen dynamics and losses in a layered agricultural soil. *Agric Ecosyst. Environ.* **18**, 333–356.

Martens, H. & Naes, T. 1989 *Multivariate Calibration.* Wiley, Chichester.

Quinn, P., Anthony, S. & Lord, E. 1999 Basin scale nitrate simulation using a minimum information requirement approach. In: *Water Quality—Processes And Policy* (ed. Trudgill, S. T., Walling, D. E. & Webb, B. W.). Wiley, Chichester, pp. 101–117.

Stone, M. & Brooks, R. J. 1990 Continuum regression: cross validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression (with discussion). *J. R. Statist. Soc. B* **52**, 237–269; corrigendum **54**, 906–907.

Sundberg, R. 1993 Continuum regression and ridge-regression. *J. R. Statist. Soc. B* **55** (3), 653–659.

Wold, H. 1975 Soft modelling by latent variables; the non-linear iterative partial least squares approach. In: *Perspectives in Probability and Statistics, Papers in Honour of M. S. Bartlett* (ed. Gani, J.). Academic Press, London, pp. 114–142

Young, P., Parkinson, S. & Lees, M. 1996 Simplicity out of complexity in environmental modelling: Occam's razor revisited. *J. Appl. Stat.* **23**, 165–210.