Identifying dynamical models of nitrate propagation in agricultural drinking water: how can we help agronomists?

Vincent Laurain, Marion Gilson, Marc Benoît

Identifying dynamical models of nitrate propagation in agricultural drinking water: how can we help agronomists?

V. Laurain [*,**] **M. Gilson** [*,**] **M. Benoît** [***]

[*] *Université de Lorraine, CRAN, UMR 7039, 2 rue Jean Lamour, 54519 Vandoeuvre-lès-Nancy Cedex, France.*
[**] *CNRS, CRAN, UMR 7039, France.* vincent.laurain@cran.uhp-nancy.fr
[***] *INRA SAD UR055 Aster. Mirecourt, France.* Marc.Benoit@mirecourt.inra.fr

Abstract: Since the 50 last years, the rapid development of modern agriculture in industrialised countries has considerably affected the quality of water resources, up to the point to jeopardise the capacity of rural territories to produce drinking water. Hence, agronomy has been interested in the complex nitrate biogeochemical interactions for a long time. While agronomists are able to produce very accurate physical models of nitrate propagation at different scales, their tools have a limited relevance if the information regarding geology or agriculture is missing. Consequently, in many cases, it prevents the specialists of being affirmative about the prediction of their current actions on the water quality. By opposition, a system identification methodology is here presented to predict nitrate concentration in water. It has the advantage of being applicable even when very little knowledge is available. It will be shown how external variables such as rainfall and temperature can play an important role in modelling water pollution systems. The efficiency of the approach, both in terms of prediction and physical insight, is discussed on a real life dataset.

1. INTRODUCTION

Agriculture is challenged by large scale issues, like impacts of land system changes on the preservation of environmental resources, urging agronomy to evolve.

In European Union, the WFD (Water Framework directive) is built on a strict basis: water policy is a result based policy. So, States and Agencies have to maintain water in a good state, link to chemical norms and dates to obtain these results [EC2, 2000]. So, it has become compulsory to deal with two main parameters to help decision makers in this domain: the evolution of concentration and the level of chemical contains at a precise dates. This work is hence dedicated to water quality depletion or improvement.

During the past decades many different models have been proposed in order to analyse the complex biogeochemical behaviour of nitrate (N) in agricultural soils. In [Manzoni and Porporato, 2009], 250 different models are classified in terms of mathematical features such as spatial and temporal scale or isotropy approximations. These models take into account different phenomena (denitrification, biomass growth and decay, water flux ...) and therefore require the tuning of a large number of parameters. They also require quite a large number of input such as the type of culture, the N density at different depths or the soil type [Bacsi and Zemankovics, 1995, de Willigen and Neeteson, 1985]. Hence, they are mostly exclusively validated on dedicated experimental parcels [Cavero et al., 1999, Bacsi and Zemankovics, 1995, de Willigen and Neeteson, 1985], where each required information is available. The strength of those models is their deep physical insight and the respect of a modelling protocol allowing their generalisation to other parcels.

Nevertheless, their main drawback is their inability to be tuned on parcels where some of the required knowledge is unavailable: in this case, some assumptions are required, which can average favourably at large scales, but that are hardly verifiable at smaller scales such as catchment or parcels scale[Del Grosso et al., 2006]. This problem was early acknowledged in [Ledoux et al., 2007, Beven, 2000] and considerably limits a possible dynamical analysis. In the presented application, the only available information is the compulsory N concentration measure in drinkable water sources. All the physical knowledge is missing: *e.g.* the depth of these sources, the surface of drained water, the flow, the culture types or the soil type. Estimating a model of the N propagation in the water becomes a challenging issue in this context which actually represents one of the most realistic scenario. It means that in such a situation the only analysis left is the trend static analysis: is the N concentration raising or decreasing?

All the available nitrate cycle models can be referred to as *so-called* "bottom-up" models : from the physics to the data. A diametrically opposed philosophy also emerged in the environmental field: "top-down" approach also referred to as data-based mechanistic [Young and Garnier,

[1] The authors would like to thank Gilles Rouyer (Aster unit) for the water analysis, ZAM (Zone Atelier du Bassin de la Moselle) and RésEAu Lor-Lux for their support. This project was supported by the CNRS PEPS project ContamiNit.

2006]. The model is determined using the measured data and these approaches link directly to the field of system identification. There are many environmental fields where system identification was successfully used, and one of the most prosperous field related to the presented application is the rainfall/runoff modelling ([Young and Garnier, 2006, Laurain et al., 2010, Lorent and Gevers, 1974]). This paper main contribution is to propose a data-based model for the problem of nitrate propagation modelling and is consequently mainly applicative.

The identification problem is challenging: Firstly, agronomists are only interested in physically interpretable models. Therefore, once a model is designed, the validation process can only be based on physical propositions. Secondly, the input represented by the N mass spread by farmers is unknown. Hence, it is firstly required to find some external variable correlated with N sources.

The paper is organised as follows. In Section 2, the first step of finding input through correlation analysis will be carefully explained. Based on the data mining considerations, the identification problem and the proposed model are detailed in Section 3. Finally, the results are exposed and analysed from the control theory viewpoint in 4 and from agronomic viewpoint in Section 5. Conclusions and some future directions of research are given in Section 6.

2. TEMPERATURE AND RAINFALL AS INPUT?

In this case study, the available data consists of the N concentration C^{N} in 6 underground water sources (S_1 to S_6) under farming management, located in Lorraine, Plateau Lorrain Region, France. The sample period is irregular in some parts of the data and the minimum sampling period is 15 days from the January 1st, 1990 until 2003. An example is shown in Figure 1(a).

In the presented application, system identification is a delicate problem as unlike pesticides, nitrate have many different sources which can be both natural (plant own production, cow rejections) and human (mineral fertilisation, polluted rainfall). For example, it is not unusual to find nitrate in forest underground waters without any human activity. Most unfortunately, in many realistic cases, none of the N sources measurements is available at the watershed scale level.

While human N sources cannot likely be correlated with any kind of usually measured signals, natural N sources might find possible correlated signal candidates: it is fair, for example to assume that the vegetation density is related in some way to the temperature and the rainfall. Moreover, those measures are widely and commonly available. In this study the rainfall $r(t)$ daily sampled is available and exposed in Figure 1(c). Moreover, the average daily temperature $\tau(t)$ at Nancy station, France has been downloaded from the European Climate Assessment website (Source 741) [Tank and Coauthors, 2002] and is displayed in Figure 1(b).

All the accessible data are not homogeneous in sampling frequency. In order to harmonise the sampling period between al signals, it has been chosen to interpolate C^{N} daily in order to keep most available knowledge on temperature and rainfall. Here linear interpolation has been performed. Since there isn't any available knowledge on the inter-sample behaviour, and since the aim of this study is to define a possible predictive model structure, the interpolation choice and effects are not discussed in this reduced conference format.

Before identifying any dynamic model, a possible correlation between the temperature/rainfall and the nitrate concentration is studied in order to determine whether they can be considered as possible input signals. Since, the true relationship can possibly be time-varying or nonlinear, a local correlation analysis on a sliding window is driven. At each time t_i, the local correlation scores $\rho^w_{\tau,C^{\mathrm{N}}}(t_i)$ and $\rho^w_{r,C^{\mathrm{N}}}(t_i)$ defined as:

$$\rho^w_{X,Y}(t_i) = \rho(X,Y)|_{k\in\{i-w,i+w\}}, \qquad (1)$$

are computed, where w represents the sliding window size and $\rho(X,Y)$ is the correlation score between signals X and Y. Due to space restriction only the result for 3 sources S_1, S_2 and S_3 are exposed in Figure 2 for $w = 365$ (each window represents two years).

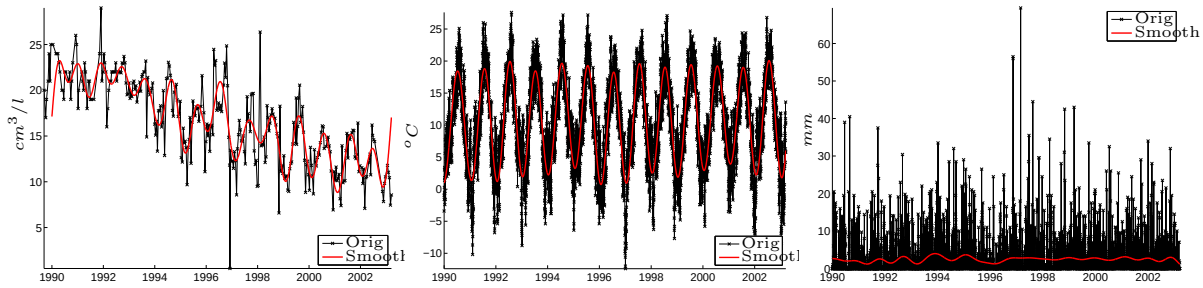
A strong correlation between temperature and N concentration appears ($\rho^{365}_{\tau,C^{\mathrm{N}}}$ reaches 0.6 on some of the sources). Similar results are obtained on the 6 sources. Since, the temperature curve is almost periodic, this highlights a pseudo-periodic behaviour of the concentration.

Correlation with the rainfall is however doubtful. Nonetheless, since concentration was originally sampled fortnightly, the signal does not contain any high frequencies, unlike the rainfall signal. Therefore, before excluding the rainfall possible contribution, the same correlation study is driven after a low pass filtering of the signals.

In the sequel, a signal $x_T(t)$ defines the low pass filtered version of $x(t)$ with a filter cut-off frequency of $1/T$ (1/days). For example $\tau_{200}(t), C^N_{200}(t)$ and $r_{200}(t)$ are displayed in Figure 1. It must be noticed that during this phase, the choice $T_{\mathrm{o}} = 200$ days is not critical as only the existence of a correlation is evaluated. The associated local temperature/concentration and rainfall/concentration correlation scores $\rho^{365}_{\tau_{200},C^{\mathrm{N}}_{200}}$ and $\rho^{365}_{r_{200},C^{\mathrm{N}}_{200}}$ are computed as defined in(1). The results are exposed in Figure 3 for both the temperature (red curves) and the rainfall (blue curves). In this Figure, it can be clearly be concluded that the low frequency rainfall components are also correlated to the low frequency components of the N concentration.

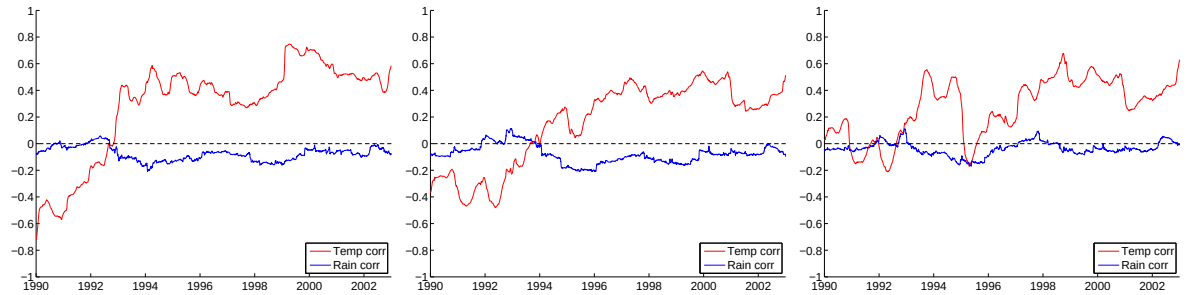
Here is an important preliminary conclusion for system identifiers. In many environmental applications, important measures are missing due to high costs or complex data acquisition processes. In those cases, temperature and rainfall should not be overlooked as an important source of information. For example, it is clear how temperature and rainfall play a major role in the water cycle. Moreover, temperature is a good indicator of agricultural practices and/or vegetation level. Hence, even in pollution applications, rainfall and temperature might be strongly linked to the system under study.

Finally, it can be noticed from Figure 3 that the correlation varies much over time. This must be carefully taken in account since a time-varying behaviour seriously compli-



(a) Nitrogen concentration in underground water source S_3 *(b)* Temperature *(c)* Rainfall

Fig. 1. Raw and smoothed data



(a) Source 1 *(b)* Source 2 *(c)* Source 3

Fig. 2. Correlation between the temperature/rainfall and the N concentration using the raw data of S_1,S_2 and S_3.

cates the identification problem. Indeed, the number of data points (approximately 300 for each source) is hardly enough to identify complex time-varying behaviours.

The next section defines the identification problem once the rainfall and temperature have been defined as input.

3. IDENTIFICATION PROBLEM STATEMENT

As in any data-based modelling procedure, an identification dataset is needed in order to optimise a model and a validation dataset is required to cross-validate the obtained model. Thirteen years of data is available with 15 days sampling period. It means that approximately 315 data points are measured. Under these conditions, splitting the dataset into validation and estimation dataset means that at most, a 160 points dataset is available for identification. Under these conditions, it is hardly achievable to define a strongly nonlinear structure or time-varying structure. Hence, the model structure will be restricted to linear models only.

In order to avoid a strongly changing behaviour in the identification set, years from 1990 to 1994 are avoided since they exhibit the most drastic change in local correlation (see Figure 3). Moreover, 1996 and 1997 are according to agronomists the driest and rainiest years respectively by far. Hence, years 1996 to 1999 should englobe the most marginal as well as the most average behaviours and are retained for identification while the validation dataset is taken as the whole data set.

Furthermore, the drift on C^N (appearing on the original N concentration in Figure 1(a)) is removed for identification purposes. This is performed by removing low-pass filtered signals and therefore defining the following centred signals:

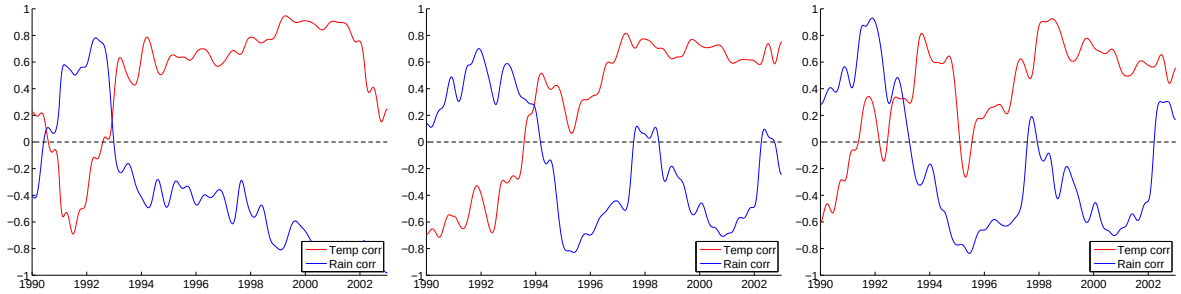
$$\begin{cases} C_c^{\mathrm{N}}(t) = C^{\mathrm{N}}(t) - C_{1460}^{\mathrm{N}}(t), \\ \tau_c(t) = \tau(t) - \tau_{1460}(t) \\ r_c(t) = r(t) - r_{1460}(t). \end{cases}$$

It must be noticed that instead of a low-pass filtered signals, second order polynomial curve could be chosen to represent the drift, with same performances.

Finally, the identification problem can be stated as:
Given the nitrate concentration data (output signal) $C_c^{\mathrm{N}}(t)$, the rainfall data $r_c(t)$ and the temperature data $\tau_c(t)$ sampled at time t_k $k = 1..N$, estimate a linear model representing the N concentration propagation relationship for the given parcel.
The model class considered in this case study can be described as the following continuous-time Multi Input Single Output (MISO) Output Error (OE) hybrid model:

$$\mathcal{M} \begin{cases} \mathring{C}_c^N(t) = \dfrac{\sum_{j=0}^{m_r} b_j p^j}{s_r^n + \sum_{j=1}^{n_r-1} a_j p^j} r_c(t) + \dfrac{\sum_{i=0}^{m_\tau} \beta_i p^i}{s^{n_\tau} + \sum_{i=1}^{n_\tau-1} \alpha_j p^i} \tau_c(t) \\ C_c^N(t_k) = \mathring{C}_c^N(t_k) + e(t_k), \end{cases} \quad (2)$$

where n_r,n_τ and m_r,m_τ are the unknown model orders, p is the differentiation operator and $\{\alpha_i, \beta_i, a_i, b_j\}$ are the model parameters to be estimated. In the presented model, the noise term $e(t_k)$ is assumed to be a white noise stochastic process. Naturally, the identification method used in order to fit the model plays a crucial role, especially in cases where the amount of noise is consequent. Various estimation methods are available in the literature for continuous-time models [Garnier and Wang (Editors), March 2008]. For this application, the so-called simplified



(a) Source 1 *(b)* Source 2 *(c)* Source 3

Fig. 3. Correlation between the nitrate concentration, the temperature and the rainfall in the low frequency domain

refined instrumental variable algorithm (SRIVC) [Young and Jakeman, 1980] is used. From the theoretical viewpoint, this identification method exhibits robustness with regards to high or uncommon noise conditions (unbiased models) and in case of system approximation, it has also demonstrated good results in environmental modelling [Laurain et al., 2010, Young and Garnier, 2006].

4. RESULTS

In order to determine the model quality, the model fit quality is assessed using the coefficient of determination defined as:

$$R^2 = 1 - \frac{||\hat{C}_c^{\mathrm{N}}(t_k) - C_c^{\mathrm{N}}(t_k)||^2}{\|\hat{C}_c^{\mathrm{N}}(t_k) - \bar{C}_c^{\mathrm{N}}\|^2}, \tag{3}$$

where $\hat{C}_c^{\mathrm{N}}(t_k)$ is the output simulated from the identified model on the validation data set and $\bar{C}_c^{\mathrm{N}}$ is the average value of C_c^{N}. It must be noticed that $R^2 = 1$ means perfect fit, $R^2 = 0$ means that the simulated output is only as predictive as the output average, while $R^2 < 0$ means that the simulated output is not predictive. Usually, $R^2 < -1$ is considered as a failure to find a suitable model.

Usually, to determine orders n_r,n_τ and m_r or m_τ in (2) criteria taking into account the parsimony of the model such as Akaike Information Criteria are used. Nevertheless, in the presented application, only $n_r = n_\tau = 1$ and $m_r = m_\tau = 0$ give some predictive models and therefore the most parsimonious structure is directly retained:

$$\mathcal{M} \begin{cases} \mathring{C}_c^N(t) = \dfrac{b_0}{p+a_1} r_c(t) + \dfrac{\beta_0}{p+\alpha_1} \tau_c(t), \\ C_c^N(t_k) = \mathring{C}_c^N(t_k) + e(t_k) \end{cases} \tag{4}$$

Figure 4 exposes the measured concentration (grey), the model output (black) as well as the dataset part used for identification (green). It can be noticed from Figure 4 that even though the input are external variables, a general good fit is observed for all sources. Another observable fact is that, as expected from Figure 3, the system behaviour seems to vary over time as expected from the correlation analysis. However, a striking fact is that the model is able to fit quite precisely for years 1994 to 2003, but is unable to reproduce the behaviour of years 1990 to 1994. Actually, no model could be suitably estimated from dataset 1990-1994 in this study. For all the sources, the system seems to undergo a radical behavioural change around 1994, very abrupt in S_1, S_2,S_3 and S_4, even if 1994 was not in the estimation set.

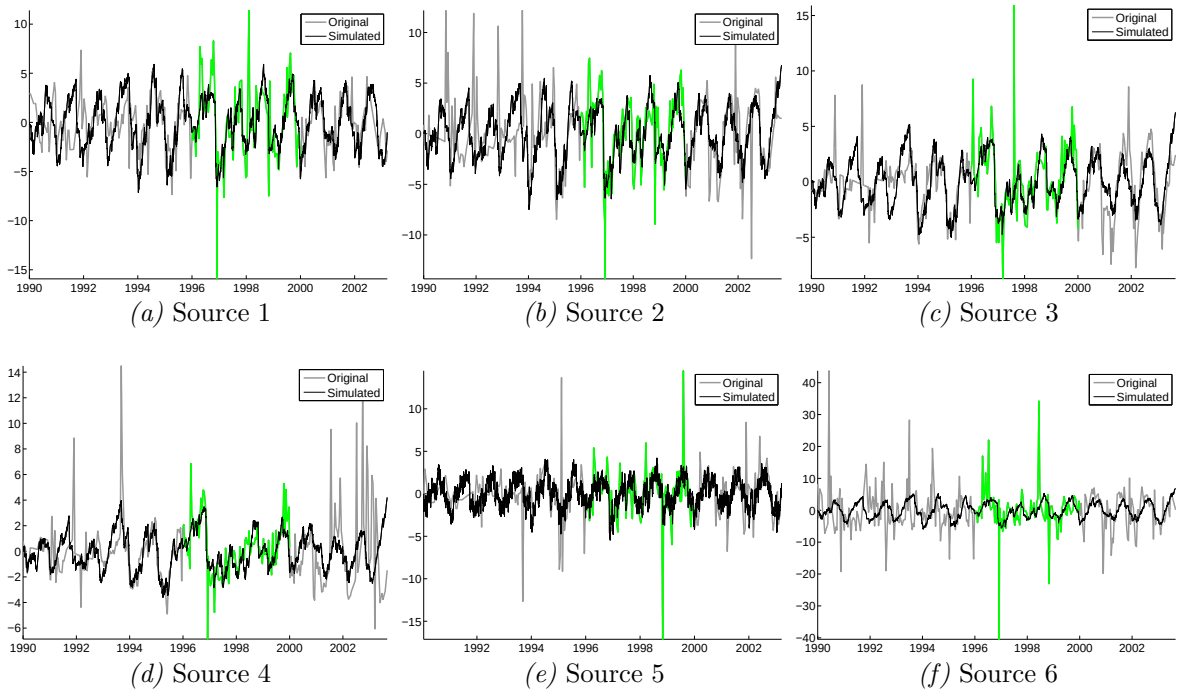
In order to properly assess how and when behavioural changes occur, a local coefficient of determination is computed at each point in a local window of two years $R(t_i) = R^2|_{k\in\{i-365,i+365\}}$: the results are exposed in Figure 5 in black. For all sources, the coefficient of determination shows an explicit transition between two behaviours from an unpredicted zone in the first part of the data, to a predictive behaviour towards the end. This jump is very abrupt for S_1 to S_5 and takes place at the beginning of 1994 as shown by the coefficients of determination which reaches values up to 0.7. This value can be considered as low in a usual system identification framework. Nevertheless, in the present context, without any explicit input of N, sampling difficulties and low number of data, this fitting score can be considered as a good fit.

Moreover, as previously pointed out, the correlation between the inputs and output was much better in the low frequency domain. Therefore, Figure 5 also exposes the coefficient of determination between $\hat{C}_{c_{200}}$ and $C_{c_{200}}$ in red. It shows that in the low frequencies domain, the coefficient of determination reaches up to 0.9 which is an undeniable fit and indicates that the yearly trend of nitrate in those water is nearly completely predictable from the rainfall and the temperature but only from 1994 on. This statement is however mild for S_4 which shows unpredicted high frequency signals from 2001 on.

ince the identified model does not represent directly a physical relationship, the absolute value of the parameters is probably of little interest. Nonetheless, it appears that all transfer functions linked to the temperature have a positive static gain while all rainfall transfer function have a negative static gain.

The model has been proposed from the data and according to data-based mechanistic principles, it can therefore only be validated through physical facts. Therefore the main outcome of this study is to propose from this black-box approach physically interpretable facts which are here:

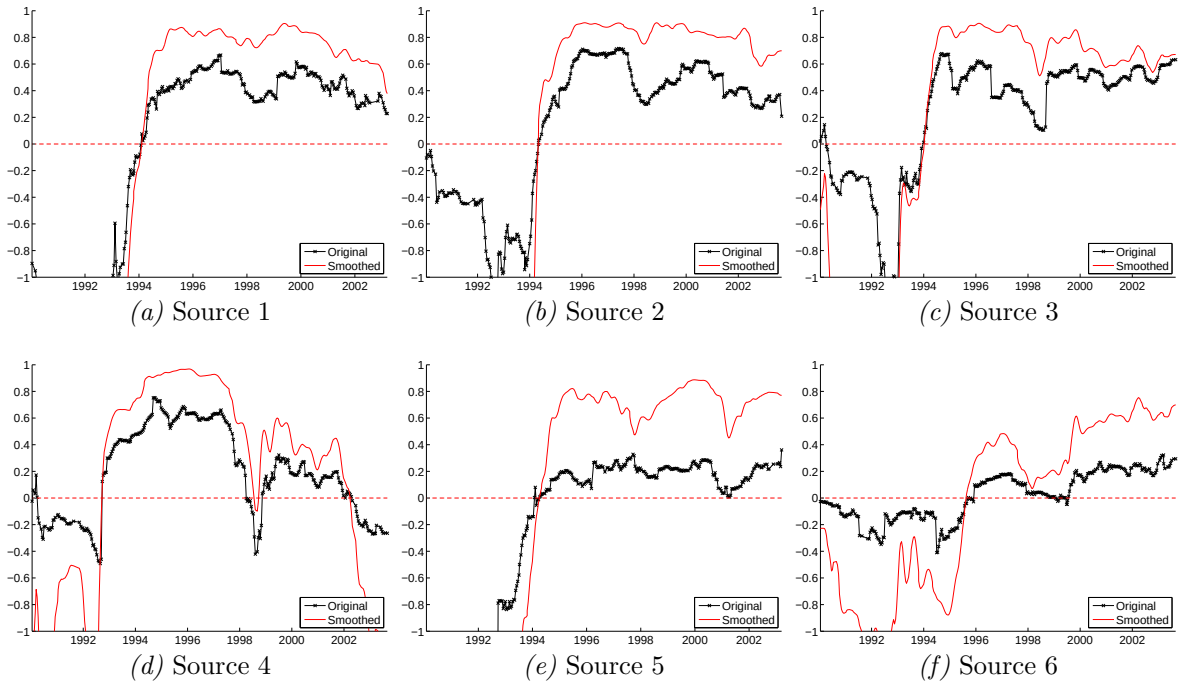
- C1 The rainfall has a negative effect on the nitrate concentration.
- C2 The temperature has a positive contribution on the nitrate concentration.
- C3 The studied sources undergoes a major behavioural change in year 1994.



(a) Source 1 *(b)* Source 2 *(c)* Source 3

(d) Source 4 *(e)* Source 5 *(f)* Source 6

Fig. 4. Measured concentration and simulated model concentration



(a) Source 1 *(b)* Source 2 *(c)* Source 3

(d) Source 4 *(e)* Source 5 *(f)* Source 6

Fig. 5. Coefficient of determination between predicted and measured output

5. THE AGRONOMIST VALIDATION

In agronomy, the temperature is a good indicator of vegetation growth season. In these temperate zones, the $6°C$ is the basis of two main phenomena: vegetation growth, and nitrate mineralization in the soil. Hence, temperature can be used as an indicator of natural nitrate production which validates C2. Usually, rainfall minus evapotranspiration is a good indicator of nitrate movement in the soils and in the aquifer: without rainfall, N are stable in the soils, with low rainfall, they are moving, because very soluble, through the soils to the aquifers raising the N concentration. Nevertheless, above a critical amount of rain, the nitrate does not interact with water anymore, decreasing their concentration in the water, which validates C1.

Finally, until 1992, each parcel was managed independently without any common fertilising policy. However, since 1993 (with an actual application in 1994), a local water resource management operation has taken place and all farmers have been asked to strongly reduce the amount of nitrate spreading. For all studied sources, it seems that the simple proposed model is able to clearly emphasise this change which validates C3. It is even more striking for Sources 1 and 2. The physics of these sources are very similar (slope, size, location). Nevertheless, before 1994, the concentration signals are strongly different (their drift is actually opposite). Nevertheless, the identified models are nearly exactly equal and their fit similar. Hence, after the spreading has become homogeneous, the model clearly catches the similarities of these parcels. Furthermore, the strong reduction of fertilisers input from 1993 encourages us to think that the behaviour after 1994 is close to a natural behaviour. In natural parcels, the nitrate is mainly issued from the soil organic matter mineralization by bacterial activities which is strongly correlated to the temperature.

6. CONCLUSIONS

The problem of nitrate propagation in water has been addressed in this paper. Most agronomists mechanistic models fail to determine the future water quality since in practice, much physical and costly knowledge is missing. In this paper, a data-driven protocol has been detailed in order to propose dynamical models of the nitrate propagation in drinking water under agricultural soils, at watershed scale. A correlation study has shown that temperature and rainfall are strongly linked to the nitrate concentration in water. The causality has even been justified by agronomists. A simple first order dynamical model has been identified which has shown extremely good prediction capabilities on the years following a major agricultural practice change dedicated to minimise the nitrate loads in drinking water. From this black-box approach, some physical propositions have been derived which could all be validated from agronomical knowledge. This emphasises that despite the data-driven nature of the approach, the proposed model seems to well represent physical behaviour. Finally, the far end goal for such model is the concentration drift prediction which could be a very slow dynamic. Some further work is hence needed in order to deeper investigate how carefully each farmer has followed the spreading advice and for how long. Should it be correlated to the proposed model fit, it will be investigated how much the model fit can predict the slow trends which are so far not identifiable from the data and without the knowledge of spread nitrates amount.

REFERENCES

Directive 2000/60/EC of the European Parliament and of the council of 23 October, 2000 establishing a framework for community action in the field of water policy. Official J Eur Commun, 2000.

Z. Bacsi and F. Zemankovics. Validation: an objective or a tool? results on a winter wheat simulation model application. *Ecological Modelling*, 81:251–263, 1995.

K. J. Beven. Uniqueness of place and process representations in hydrological modelling. *Hydrology and earth system sciences*, 4:203–213, 2000.

J. Cavero, R.E. Plant, C. Shennan, D.B. Friedman, J.R. Williams, J.R. Kiniry, and V.W. Benson. Modeling nitrogen cycling in tomato-safflower and tomato-wheat rotations. *Agricultural systems*, 60:123–135, 1999.

P. de Willigen and JJ. Neeteson. Comparison of six simulation models for the nitrogen cycle in the soil. *Fertilizer research*, 8, 1985.

S.J. Del Grosso, W.J. Parton, A.R. Mosier, M.K. Walsh, D.S. Ojima, and P.E. Thornton. Daycent national-scale simulations of nitrous oxide emissions from cropped soils in the united states. *Journal of Environmental quality*, 35:1451–1460, 2006.

H. Garnier and L. Wang (Editors). *Identification of Continuous-time Models from Sampled Data*. Springer-Verlag, London, March 2008.

V. Laurain, M. Gilson, S. Payraudeau, C. Grégoire, and H. Garnier. A new data-based modelling method for identifying parsimonious nonlinear rainfall/flow models. In *In proceedings of the International Congress on Environmental Modelling and Software (IEMSS 2010)*, Ottawa, Ontario, Canada, July 2010.

E. Ledoux, E. Gomez, J. M. Monget, C. Viavattene, P. Viennot, A. Ducharne, M. Benoit, C. Mignolet, C. Schott, and B. Mary. Agriculture and groundwater nitrate contamination in the seine basin. the sticsmodcou modelling chain. *Science of the total environment*, 375:33–47, 2007.

B. Lorent and M. Gevers. Identification of rainfall/runoff processes. *Proceedings of the 4th IFAC Symposium on Identification and parameter Estimation*, pages 735–744, 1974.

S. Manzoni and A. Porporato. Soil carbon and nitrogen mineralization: Theory and models across scales. *Soil Biology & Biochemistry*, 41:1355–1379, 2009.

A.M.G. Klein Tank and Coauthors. Daily dataset of 20th-century surface air temperature and precipitation series for the european climate assessment. *Journal of Climatology*, 22:1441–1453. Data and metadata available at `http://www.ecad.eu`, 2002.

P. C. Young and H. Garnier. Identification and estimation of continuous-time, data-based mechanistic (DBM) models for environmental systems. *Environmental Modelling & Software*, 21, Issue 8:1055–1072, August 2006.

P. C. Young and A. Jakeman. Refined instrumental variable methods of recursive time-series analysis - part III. extensions. *International Journal of Control*, 31, Issue 4:741–764, 1980.