Data-driven modeling for water resource quality over long term trends

Vincent Laurain, Marion Gilson, Marc Benoît

International Environmental Modelling and Software Society (iEMSs)
7th International Congress on Environmental Modelling and Software
San Diego, California, USA, D.P. Ames, N. Quinn (Eds.)
http://www.iemss.org/society/index.php/iemss-2014-proceedings

Data-driven modeling for water resource quality over long term trends

V. Laurain[a], M. Gilson[a], M. Benoît[b]

[a] *Université de Lorraine, CRAN, UMR 70239, 2, avenue de la forêt de Haye, Vandoeuvre-lès-Nancy Cedex, 54516, France (*vincent.laurain@univ-lorraine.fr*)*
CNRS,CRAN, UMR 7039, France

[b]*INRA SAD UR055 Aster, 662 avenue Louis Buffet F-88500 Mirecourt*

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, there has been interest within the field of agronomy to study complex nitrogen biogeochemical interactions over a long time period. Nevertheless, if the agronomists are able to produce very accurate models at different scales, they have a limited number of available tools in order to cope with quality measurements in water sources for which very little information about the geological information is available. It prevents the specialists of being affirmative about the prediction of their current actions on the water quality. By opposition, system identification can deliver dynamical models from measured data: they cannot be generalised but offer strong insight without any *a priori*. This applicative paper introduces a data-driven model software for both modelling the nitrogen propagation in drinking water and offering new decision tools to stakeholders.

Keywords: data-driven modelling, nitrogen propagation, drinking water quality.

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. Landscape agronomy has been proposed as new perspective to address these issues [Benoît et al., 2012]. In European Union, the WFD (Water Framework directive) is built on a strict basis: water policy is a result based policy. Therefore, States and Agencies have to maintain water in a good state, link to chemical norms and dates to obtain these results [2000/60/EC, 2000]. It has hence become compulsory to deal with two main parameters to help decision makers in this domain: the evolution of concentration and the level of chemical contents at precise dates (2015, 2021, 2025). 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 Nitrogen (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 (cereals,vegetables...), 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 from measurements. 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 cannot be applied at smaller scales such as catchment or parcels scales [Del Grosso et al., 2006]. In the presented application, the only available information is the compulsory N concentration measure in drinkable water sources. There is not any

knowledge about the depth of these sources, the surface of water they drain, their flow or the soil type. Modelling and understanding the N propagation into the water is a challenging issue in such a situation which actually represents the most realistic scenario in most of drinkable sources.

In opposition to the traditional approaches, a "top-down" approach is proposed in this paper: the model is determined using the measured data and despite some slight vocabulary divergences, these approaches link directly to the field of system identification. The field of system identification uses statistical methods to build mathematical models of dynamical systems from measured data. Their 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 [Laurain et al., 2010; Young and Beven, 1994].

This paper main contribution is to propose a data-driven model, along with its physical validation process. The obtained model aims at bridging the gap between the modelling possibilities with regards to slow/complex dynamics effects and the stakeholders expectations. It is organised as follows. In Section 2, the issues and dead-end of traditional modelling and nitrogen propagation are detailed. In Section 3, the proposed data-driven model is explained along with some algorithmic aspects. Finally, the results are exposed and validated using physical principles in Section 4.1. Conclusions and some future directions of research are given in Section 5.

2 PROBLEM FORMULATION

The 2000 European Directive, the Water Framework Directive (WFD), proposed three new articles: preservation of water bodies as a whole (taking into account non-point pollution instead of only point-source pollution), an imposed schedule, and objectives defining quantified results aiming for the ecological restoration of the environment. This text is complex (because it includes several types of regulatory tools), ambitious, and is a cornerstone of the European Union's environmental policy. However, its application is delicate for a number of countries [Dworak et al., 2009] in which achieving consistency within the law has followed other pathways or which do not know how to achieve this result. France partially conformed to this directive only 6 years later through its Law on Water and Aquatic Environments [Dworak et al., 2009] where for the first time in French law, the notion of non-point pollution appeared.

In 1990, nitrate peaks reaching 70 mgNO3-/L (exceeding the European drinking standard of 50 mgNO3-/l) caused a lively debate on management strategies to deal with this N contamination. The public service delegated for water proposed a strategy to protect all of the sources used for drinking water. With the European directive in mind, a collective action called "Ferti-Mieux" started on the sources of the Haut-Saintois plateau (800 ha= 8 km^2), Lorraine, France. A list of agricultural new practices has been proposed to farmers and many of them (85 % of cultivated surface) adopted these improvements to protect water resources. The main changes are i) farmyard management through low amount of compost (less than 20 t/ha/y) and ii) mineral fertilisation control taking into account soil mineralization (the global decreasing of input is 60 kg N/ha/y). These improvements are near from other European agricultural improvement operations for water quality protection [Kunkel et al., 2010; Lam et al., 2011].

Since then, six sources are monitored by measuring the N concentration at a fortnight rhythm during 20 years by technicians of Inra research unit (Lionel Caudy, Gilles Rouyer and Damien Foissy). The samplings have been realised directly in the source catchment, with all the legal security processes. The location of these sources is characterized by an agricultural landscape. The land use is: Forests (12 %), small rural roads (4 %) and agricultural fields (84 %). Hence, the pollution sources are only diffuse pollution one, without any building and houses in the watersheds. In this paper, due to space restrictions, we focus on two of these sources which nitrate ion concentration $C^N(t)$ is displayed in Figure 1 in grey. It must be noticed that the same modelling strategy with similar results was applied on all the available data. The goal is to study the link between the practice changes and the water quality. As there is not any available model of water dynamics in this karstic zone, and no N concentration measure in the soil is available, we propose a data-driven modelling strategy.

3 THE PROPOSED TOOL : DATA-DRIVEN MODEL

3.1 The physical-based modelling problem

In practice, long-time trend forecasting and decision help requires a dynamical model able to simulate the concentration in the water with respect to some given scenario. Nevertheless, in the presented problem, no model is available to link the agricultural N mass to the actual N concentra-

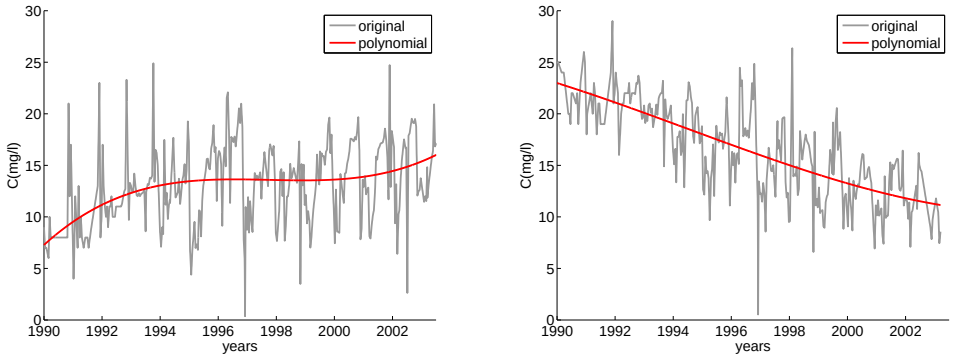

Figure 1. Measured nitrogen concentration $C_1^N(t)$ and $C_2^N(t)$ in two similar sources S_1 and S_2 along with a 3^{rd} order polynomial approximation $M_1^N(t)$ and$M_2^N(t)$

tion measured in the water. Should there be one, it would also not be possible to tune it in order to fit the current measured data : indeed, none of the N input have been measured. Hence, the commonly used modelling technique for this problem is to use a simplistic polynomial fit such as exposed in Figures 1 in red. Unfortunately, such a static model indicates the current trend but is not suitable for future forecasting. Finally and most importantly, despite the similar region, soil and agricultural policy, it appears that the polynomial curves indicate opposite results for these two sources: no conclusion can be drawn from this measured data.

3.2 The data-driven modelling problem

The solution proposed in this paper relies on the identification of a model under its transfer function form which offers the ability to interpret a posteriori the data-driven model in physical or ecological terms: this process is the *so-called* data-driven mechanistic modelling [Young and Beven, 1994]. The main advantages in using such a modelling paradigm are:

- A simplified model is obtained without requiring any physical assumption or model reduction. The obtained model might not even be possibly derived from physical assumptions. Nonetheless, only a physics-based validation and its capacity to predict physical phenomena legitimate its validity.
- If a very simple model is obtained, it also means that a source can be characterised using only a small set of parameters, helping a lot in terms of data compression in big data contexts.
- Finally, even though the obtained model cannot be directly applied to other sources without measures, the data-driven mechanistic process can be applied directly to any kind of relationship (rainfall/runoff...), at any scale as soon as some measures are available.

Regarding the specific nitrogen propagation problem, from a system theory viewpoint, two main issues have to be dealt with and are not detailed here due to space restrictions:

1 The lack of nitrogen input: none of the nitrogen soil concentration is measured. Consequently, a preliminary data-mining phase has been realised. It is based on a simple correlation analysis and the result is a possible correlation between $C^N(t)$ and both the raw rainfall $R(t)(mm/day)$ and the temperature $T(t)$ (^oC) (downloaded from the European Climate Assessment website, Source 741, [Tank and Coauthors, 2002].

2 System identification is dedicated to dynamical models: the drift represented by the polynomial approximation $M^N(t)$ varies too slowly to be identified from a reduced set of data and needs to be removed. Hence, the considered output is $C_c^N(t) = C^N(t) - M^N(t)$.

The proposed model which is the main contribution of this paper is described by the following so-called Output-Error model:

$$\mathcal{M}\left\{C_c^N(t) = \frac{b_1}{s+a_1}R_c(t) + \frac{\beta_1}{s+\alpha_1}T_c(t) + e(t), \right. \tag{1}$$

where s is the Laplace variable and $e(t)$ is assumed to be a white noise stochastic process. $C_c^N(t)$,$R_c(t)$ and $T_c(t)$ are the centred Nitrogen concentration, rainfall and temperature respectively, defined as $C_c^N(t) = C^N(t) - M^N(t)$, $R_c(t) = R(t) - \text{mean}(R(t))$ and $T_c(t) = T(t) - \text{mean}(T(t))$. b_1, a_1, α_1, b_1 are the parameters to be identified and they fully characterise the presented structure. The same structure was proposed from the data for all the sources analysed in this study (6 in total). Hence, this dynamical model is a simple first order model, with two inputs and one output. Please note that other more complicated models such as nonlinear Hammerstein and Wiener structures along with nonparametric approaches (Support Vector Machines) have been tested. Nevertheless, probably due to the amount of noise and the relatively low number of data points, the results were not significantly better (or worse), lead to an harder interpretation, and therefore, the simplest structure was retained.

This model implies that the temperature and the rainfall should affect consistently the nitrogen concentration in water. This statement can naturally not be derived from any physical logic. Hence, one would think that this type of model does not have any legitimacy. However, should this model be physically validated (temperature could be correlated to an actual N source), it would be very interesting in the sense that temperature and rainfall are widely available measures. Before drawing any conclusion identification must be performed. Nonetheless, identification is not a trivial matter either in such context. While the white noise assumption is fair when measurement noise is the major perturbation source, the problem is more complicated here : noise on the inputs (the rain and the temperature are not measured exactly above the considered parcel), missing inputs, system approximation... Hence, the identification needs to be robust to noise modelling errors. For this application, the so-called simplified refined instrumental variable algorithm (SRIVC) [Young and Jakeman, 1980] has been chosen. It hands out unbiased estimates in case the true noise is zero mean, independently of its distribution and it was successfully used in many other environmental applications [Laurain et al., 2010; Young and Beven, 1994]. Consequently, before presenting the results, here are all the steps necessary for the identification of the proposed N propagation model:

- Step 1: Force all the data to the same sampling period : in our application $C^N(t)$ was sampled every 15 days, and was linearly interpolated to fit the one day sampling period of the rainfall and temperature.
- Step 2: Compute the polynomial curve and subtract it to the original measured data.
- Step 3: Split the data into a identification dataset and a validation dataset.
- Step 4: Use the identification dataset to identify a linear structure using the SRIVC function (a free Matlab®version can be downloaded with the Contsid Toolbox at `http://www.cran.uhp-nancy.fr/contsid/`) using different model orders.
- Step 5: Use the validation dataset in order to choose the most suitable order (for example using Young's Information criterion [Young and Jakeman, 1980] and display the obtained model (resulting here in (1)).

The obtained results are provided in the next section.

4 MODEL VALIDATION

4.1 Physical propositions

In this section, the identification results are depicted in the form of physical propositions and data fit. Before detailing the results, it must be emphasised here that no *a priori* has been used in order to build this model and it has been solely built using commonly measured data. The identification data has been chosen as a period of 4 years between 1997 and 2001 as they contain the most different rainfall scenarii and the rest is used for validation purpose. The coefficients of the identified models for the two presented sources are exposed in table 1. Naturally, the equations are not homogenous and therefore, the absolute value of these parameters are of little interest. Nevertheless, it can be seen that the b_1 coefficients are negative for both sources (and actually all other studied sources).

Source	rainfall		temperature	
	a_1	b_1	α_1	β_1
S_1	0.016	-0.050	0.021	0.0007
S_2	0.003	-0.023	0.007	0.0002

Table 1. Identified parameters

This indicates a negative static gain on the rainfall transfer function. The exact opposite statement can be concluded for the temperature transfer function.

The simulated output of the identified model can be computed. The estimation data, validation data and simulated output are exposed in Figure 2. Two striking facts appear from the simulation results. Even though these two parcels behave differently from the trend perspective, their dynamic is actually quite similar. The identified model outputs are actually nearly identical for both of these sources. Furthermore, the model output fits quite well the original data (at least from in the estimation dataset) considering that the rainfall and temperature are not actual sources of nitrogen. In order to quantify this fit, a local score inspired from the NASH score [Nash and Stucliffe, 1970] is used and is defined as:

$$NASH(t_k) = 1 - \frac{||\hat{C}_c^{\mathrm{N}}(t_k) - C_c^{\mathrm{N}}(t_k)||^2_{k-N,k+N}}{||\hat{C}_c^{\mathrm{N}}(t_k) - \mathrm{mean}(C_c^{\mathrm{N}})||^2_{k-N,k+N}}. \tag{2}$$

where $\hat{C}_c^{\mathrm{N}}(t_k)$ is the output simulated from the identified model and the notation $||.||^2_{k-N,k+N}$ depicts the ℓ_2 norm on the samples between $k-N$ and $k+N$. Here, N is chosen as 365 days. In other words, the presented score represents the NASH score on a local 2 years long window on each time sample k. It must be noticed that $NASH = 1$ means perfect fit, $NASH = 0$ means that the simulated output is only as predictive as the output average, while $NASH < 0$ means that the simulated output is not predictive.

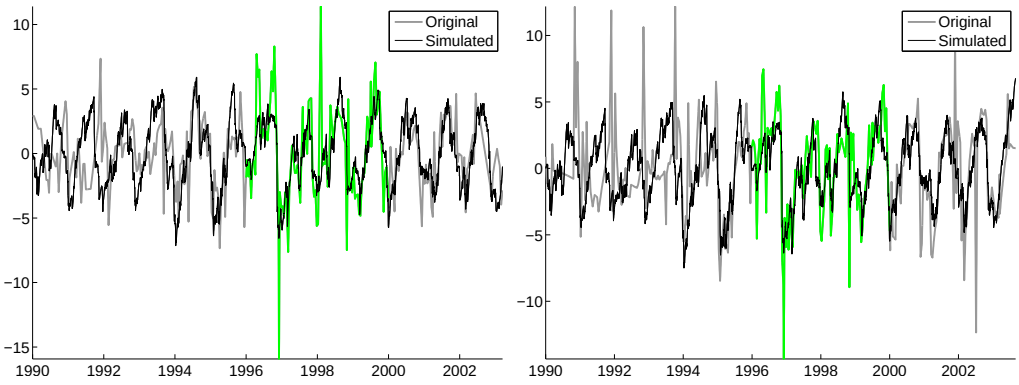

Figure 2. Original centred concentration $C_c^N(t)$ (estimation data in green and validation data in grey) along with the model output simulated from the rainfall and temperature data, for sources S_1 (left) and S_2 (right).

The fitting score quantification is displayed in Figure 3. It quantitatively appears now that both sources (actually all studied sources) evolve from a non-predictive to a predictive zone around 1994. From a system theory viewpoint, it could be argued that the obtained fit is rather poor. Nevertheless, the main unpredicted part is located in the high frequencies which is not surprising considering the relatively high sampling period taken for concentration measurements with respect to rainfall and temperature. Finally, by taking into account that the true model inputs are unknown, this models fits the data really well.

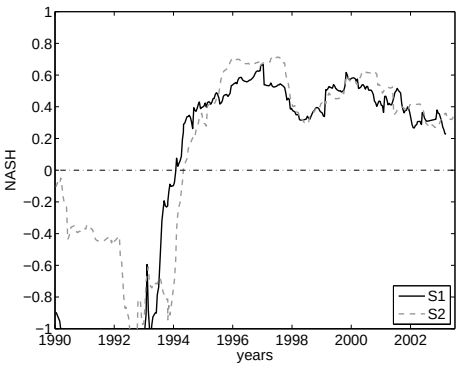

Figure 3: Local NASH score for S_1 and S_2

To sum it up, the result of the data-driven modelling are the following physical statements:

- H1: The rainfall acts negatively on the nitrogen concentration.
- H2: The temperature acts positively on the nitrogen concentration.
- H3: The dynamic behaviour of S_1 and S_2 are similar.
- H4: The sources undergo a major change around the beginning of year 1994, even though this year was not used for identification.

4.2 The physical validation

In agronomy, temperature is a good indicator of vegetation growth season. In our temperate zones, the $6^{\circ}C$ is the basis of two main phenomena: vegetation growth, and nitrate mineralization in the soil. Therefore, temperature is used as an indicator of biological activity in the fields, which is strongly involved in soil nitrification which validates H1. For agronomists, rainfall minus evapotranspiration is a pertinent indicator of nitrate movement in the soils and in the aquifer: without rainfall, N are stable in the soils while with high level of rainfall, they are moving through the soils to the aquifers. When the amount of rain is low, the N solubility being high, the concentration of N will raise in the water. On the contrary, if the amount of rain is high, once the N soil nitrification decreases under a critic mass, the N concentration will decrease. This fact strongly validates H2. Consequently, these two variables are in the heart of N leaching model and strongly correlated to N inputs [Brisson et al., 2010].

For our situation, these sources are situated very closely in the same geographic location on this karstic plateau with a similar situation in terms of climate and soil behaviours: they should behave the same (H3). Finally, the only change on these parcels is the farmers' practice changes, from a high level of inputs until 1992, to a strong decreasing in 1993 with the initiation of "Ferti-Mieux" local water resource management operation. For all studied sources, it seems that the simple proposed model is able to emphasise this change clearly in time despite the fact that this *a priori* knowledge was not used in the modelling process (H4).

Consequently, while this model could not have been possibly derived from physical reasoning, it seems that it is strongly supported by the physical validation. This model takes as an input only natural measurements. Naturally, it means that this model is unable to emphasize agricultural practice changes when most of the nitrogen source take origins from human activity. That could also mean that this model indicates how naturally behaves the source. More data would be needed in order to assess this statement. While some of the dynamics are well predicted by this model, the slow trends and the high frequencies phenomena could though not be captured. Hence, either the trends are caused by earlier data and a slower dynamic integrative effect is present or they are caused by measures taking place in the unmodelled high-frequencies phenomena. In order to study the latter assumption, a high frequency sampled-data would be needed and the current acquisition might not be sufficient to assert the future behaviour of the sources. Nevertheless, should the proposed model represent a natural behaviour, then the closer the N concentration to this model output, the more likely to see the N concentration decrease at some point. However, Further study is needed for this purpose.

5 Conclusions

In conclusion, a data-driven model has been proposed for nitrogen propagation in drinking water.It must be well understood that such an approach can be applied only on instrumented lands as the model drives the model definition. Hence, it does not require the same amount of a priori as physical based-model, but it has other limitations. For example, an identified model cannot therefore not be directly extended to other similar parcels as its parameters cannot be directly linked to any actual physical process. Nevertheless, it was shown that most of the yearly dynamics could be predicted by solely using the rainfall and temperature when the human nitrogen source is low, giving a completely different insight on the system as usual. It can be pointed out that this data was not acquired for system identification purposes and therefore, the method can be considered as applicable on commonly available data. Even though the data-driven model does not represent any physical relationship it still can be validated from physical principles. The fact that it is able to detect a change in the agricultural policy from strong nitrogen loads to considerable decreased loads could even support the fact that this model can well represent the natural behaviour of a source. The more the model fits, the more the source is natural. Nevertheless, deeper investigations are required in order to validate this assumption. Hence, the presented model could not estimate yet the long terms trends. Nevertheless, the current trends might be a residual of past effects. Should it be the case, and depending on the time constants (which might be several dozens of years) the trends will never be possibly forecasted. However, should this model represent the naturalness of a source, then it would give the stakeholders a tool to assess the naturalness of their watershed and assert whether the current trends are caused by current policies or not.

6 ACKNOWLEDGEMENTS

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

REFERENCES

2000/60/EC, D. (2000). *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.

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

Benoît, M., Rizzo, D., Marraccini, E., Moonen, A., Galli, M., Lardon, S., Rapey, H., Thenail, C., and Bonari, E. (2012). Landscape agronomy: a new field for addressing agricultural landscape dynamics. *Landscape ecology*, 27(10):1385–1394.

Brisson, N., Gate, P., Gouache, D., Charmet, G., Oury, F., and Huard, F. (2010). Why are wheat yields stagnating in europe? a comprehensive data analysis for france. *Field Crops Research*, 119-1:201–212.

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

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

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

Dworak, T., Campling, M. B. V. L. P., Kampa, E., and Thaler, M. M. R. T. (2009). Wfd and agriculture linkages at the eu level 2007-2013. *Summary report on an in-depth assessment of RD-programs*.

Kunkel, R., Kreins, P., Tetzlaff, B., and Wendland, F. (2010). Forecasting the effects of eu policy measures on the nitrate pollution of groundwater and surface waters. *Journal of Environmental Sciences*, 22-6:872877.

Lam, Q., Schmalz, B., and Fohrer, N. (2011). The impact of agricultural best management practices on water quality in a north german lowland catchment. *Environmental Monitoring and Assessment*, 183:351–379.

Laurain, V., Gilson, M., Payraudeau, S., Grégoire, C., and Garnier, H. (2010). 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.

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

Nash, J. and Stucliffe, J. (1970). River flow forecasting through conceptuel models. part 1-a discussion of principles. *Journal of Hydrology*, 10, Issue 3:282–290.

Tank, A. K. and Coauthors (2002). 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`.

Young, P. and Beven, K. (1994). Data-based mechanistic modelling and the rainfall-flow nonlinearity. *Environmetrics*, 5:335–363.

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