Metamodeling methods that incorporate qualitative variables for improved design of vegetative filter strips

Claire Lauvernet, Céline Helbert

Metamodeling methods that incorporate qualitative variables for improved design of vegetative filter strips.

Claire Lauvernet[a,*], Céline Helbert[b]

[a] *INRAE, FR-69616 Villeurbanne Cedex, France*
[b] *Univ. Lyon, UMR CNRS 5208, Ecole Centrale de Lyon - ICJ, France*

Abstract

Significant amounts of pollutant are measured in surface water, their presence due in part to the use of pesticides in agriculture. One solution to limit pesticide transfer by surface runoff is to implement vegetative filter strips. The sizing of VFSs is a major issue, with influencing factors that include local conditions (climate, soil, vegetation). The BUVARD modeling toolkit was developed to design VFSs throughout France according to these properties. This toolkit includes the numerical model VFSMOD, which quantifies dynamic effects of VFS on site-specific pesticide mitigation efficiency. In this paper, a metamodeling, or model dimension reduction, approach is proposed to ease the use of BUVARD and to help users design VFSs that are adapted to specific contexts. Three different reduced models, or surrogates, are compared: a linear model, GAM, and kriging. It is shown that kriging, implemented with a covariance kernel for a mixture of qualitative and quantitative inputs, outperforms the other metamodels. The metamodel is a way of providing a relevant first approximation to help design the pollution reduction device. In addition, it is a relevant tool to visualize the impact that lack of knowledge of some field parameters can have when performing pollution risk analysis and management.

Keywords: Metamodeling, surrogate, model reduction, kriging, Gaussian processes, qualitative inputs, categorical variables, vegetative filter strip, VFSMOD, buffer zone, pesticide, surface runoff, overland flow

1. Introduction

In recent decades, water quality degradation has become an increasing concern for society, considering its major effects on natural ecosystems and human health. In France and more generally over Europe[1], significant amounts of pollutant are measured in surface water, due in part to the use of pesticides in agriculture (Dubois and Parisse, 2017). The European Water Framework Directive advocates the development of best management practices (BMPs) to reduce pesticide transfers from the watershed to the river network. This

*Corresponding author
Email address: claire.lauvernet@irstea.fr (Claire Lauvernet)
[1] see european stastistics Eurostat : https://ec.europa.eu/eurostat/statistics-explained/index.php/Archive:Agri-environmental_indicator_-_pesticide_pollution_of_water#Context

includes implementing vegetative filter strips (VFSs, also called buffer strips or grass strips), which ensure the interception and mitigation of contaminant transfers from farm fields. VFSs are now mandatory along rivers in many countries [2], due to their recognized effectiveness in limiting surface runoff transfers of pesticides and sediments (e.g., Reichenberger et al., 2007). However, directives of this nature are regularly subject to questioning and discussion at European and national levels. For example in France, depending on the region, some rivers are classified to be protected by a VFS of 5 m length while others are not. More recently, it has been decided that an area free of pesticide treatment, of 5 m to 50 m in length depending on the chemical, should be implemented on or downslope of agricultural fields [3]. However, these regulations leave ditches and unclassified watercourses unprotected, and yet these small-scale hydrographic networks are usually the most impacted by pesticides emanating from watersheds, as they are highly exposed to drift and runoff. Whatever the regulation, scientific studies have shown that the general effectiveness of VFSs to act as a buffer can vary from 0% to 99%, depending on their design (position on the hillslope and size), and that the design of VFSs should account for agronomic conditions, soil characteristics, and climate (Muñoz-Carpena et al., 1999; Lacas et al., 2005; Dosskey et al., 2006). In this context, Carluer et al. (2017) developed BUVARD (BUffer Vegetative strip for runoff Attenuation and pesticide Retention Design tool) to design site-specific VFSs over France by simulating their efficiency in controlling surface runoff pollution as a function of local field characteristics. The BUVARD modeling toolkit combines several dynamical models (for rainfall and surface runoff entering the filter and for processes occurring within the filter) with the benchmark numerical model VFSMOD, or Vegetative Filter Strip Modeling System (Muñoz-Carpena et al., 1999; Muñoz-Carpena and Parsons, 2004; Muñoz Carpena et al., 2018; Lauvernet and Muñoz-Carpena, 2018). The method is similar to the design procedure for VFSs described in Muñoz-Carpena and Parsons (2004), but has been adapted to French conditions, in terms of input parameters and forcings given to VFSMOD. Considering local knowledge on climate, soil, cultivation practices, and water table depth, the model is run on a set of rainfall events, for several VFS lengths, and the length giving the targeted efficiency is selected for the user, including an associated uncertainty. This comprehensive method assumes that the user provides detailed field knowledge and data (such as hydrology and soil properties) that are not readily available in many practical applications, making the design procedure relatively difficult to follow.

Similar methods have been proposed for applications in the United States that present the same limits for operational purposes. A typical way to make the methodolgy more operational is to use a subset of the model simulations or to reduce the set with simple methods. Dosskey et al. (2011) developed simple relationships for sediment-bound and dissolved pollutants based on simulations of VFSMOD. Carluer et al. (2017) conducted

[2] see for example in France: `https://www.data.gouv.fr/es/datasets/eau-cours-deau-pour-la-conditionnalite/`

[3] Order of 27 December 2019 - , amending the Order of 4 May 2017 on the placing on the market and use of plant protection products and their adjuvants referred to in Article L. 253-1 of the Rural and Maritime Fisheries Code

a set of virtual scenarios with BUVARD, among which the users have to choose the most relevant considering their own situation. White and Arnold (2009), based on VFSMOD simulations, built some regressions on runoff and sediment VFS efficiency to include them in the watershed scale model SWAT. However, these regression equations involve only runoff loading and the saturated hydraulic conductivity, and thus do not properly represent the physically coupled processes occurring in a VFS as described in VFSMOD. Except for Carluer et al. (2017), these sizing methods did not take into account the presence of a shallow water table below the filter, although the water table can have a large impact on VFS efficiency with regards to water and pesticides infiltration (Muñoz Carpena et al., 2018; Lauvernet and Muñoz-Carpena, 2018; Fox et al., 2018). However, in Carluer et al. (2017), while the large number of scenarios covers a wide range of conditions, it is not possible to extrapolate to scenarios that were not simulated by the original BUVARD toolkit. Moreover, these methods do not provide uncertainty quantification, which is essential for proper use of a model for risk assessment and/or decision-making for water quality (Muñoz-Carpena et al., 2007; Shirmohammadi et al., 2006; Fu et al., 2019).

The present study aims at enabling BUVARD to be used under new climatic and agronomic conditions at a reduced computational cost with metamodeling methods (or surrogate modeling, or model reduction) that allow uncertainty to be addressed. Metamodeling is still rarely used in the water quality domain, where processes related to pesticide transfer are highly nonlinear and interacting, and lead to complex models that combine empirical and mechanistic approaches (Gatel et al., 2019a,b; Rouzies et al., 2019). Interest in metamodeling is, on the other hand, on the increase in costly environmental applications, such as calibration, data assimilation, and sensitivity analysis (Ratto et al., 2012), and in operational projects with real-time decision-making (Fienen et al., 2015). Many studies use the term metamodel to refer to a simplified model of a complex physical model, built from first principles but not based on statistical methods of automatic learning. These deterministic approaches are not considered in this paper, which focuses on statistical metamodeling. Younes et al. (2018) use a high-order polynomial chaos expansion of a flow and pesticide transport model to decrease the computational cost of Markov chain Monte Carlo calibration. The metamodel is built on reduced intervals that are obtained in a previous step by a first-order approximation method on the original model. They showed that the combined method is 70 times more efficient in time than the standard MCMC method for calibration, and that it yields accurate mean estimated values and confidence intervals. Tiktak et al. (2006) used metamodeling to develop groundwater indicators for assessing groundwater pollution risks. The method is based on analytical expression built on simulations crossing geographical zones with regression techniques. Piñeros Garcet et al. (2006) studied two metamodeling techniques applied to a nitrate leaching model: multidimensional kriging and radial based neural networks. They used the best method (kriging) to assess the probability of annual nitrate leaching concentration exceeding the legal threshold. At the European scale, Villa-Vialaneix et al. (2012) tested eight methods for surrogating another biogeochemical model, from parametric (linear model) to

non-parametric approaches. They found that random forest was the most efficient method for N_2O predictions, and support-vector machine for N leaching prediction.

The techniques surveyed above are shown to be particularly efficient for soil infiltration and groundwater processes that are slow and regular, since the deep soil that water travels through absorbs physical oscillations from rainfall. Surface processes such as runoff pesticide transfers can occur with two different types of runoff (saturation excess or infiltration excess) and depend directly on the rainfall oscillations and the soil characteristics but also on the treatment date before the rainfall event, the pollutant chemical properties, and other factors. The effects of these different phenomena are complex and difficult to model. This may explain why very few studies have proposed surrogate modeling of water and pesticide transfers taking into account surface and subsurface interactions. Adriaanse et al. (2017) generated a metamodel based on regression to determine the peak concentration of pesticides in the FOCUS surface water scenarios used in the European Union registration procedures. Regression is defined as a function of the mass concentration in the runoff water leaving the treated agricultural fields, and is based on strong simplification assumptions concerning pesticide reaction processes. In short, the methods most commonly used in water quality metamodeling are Gaussian processes (or kriging, Rasmussen, 2006) and regression. But in all these applications, the surrogate is built on quantitative variables only. However, the VFS sizing tool, BUVARD, and the physical processes it represents (water and pesticide transfer at the surface/subsurface interface) include some complex characteristics, including nonlinearities, due to the dependence on qualitative inputs (or categorical variables). Indeed, two major inputs, the type of soil of the VFS and the type of rainfall event, have been defined in BUVARD for operational purposes as substitutes for functional inputs (rainfall hyetograph) and for correlated inputs that are the hydrodynamic properties of the soil (such as saturated hydraulic conductivity, porosity, and retention curve parameters). Qualitative inputs generate discontinuities in the model response that many methods are unable to deal with, removing the smoothness of the model output, which is generally a necessary condition for building a metamodel. Yet, there is a clear need to include these disruptive qualitative inputs (Higdon, 2015; Zhang and Notz, 2015). Recently, kriging has been extended to take into account categorical inputs. Chen et al. (2013) proposed several kernels to account for categorical variables and tested these methods successfully on very simple models, based on qualitative variables only, and Roustant et al. (2018) presented a kriging-based approach with mixed categorical and continuous inputs, not limited by a large amount of qualitative values.

The objectives of this work are both methodological and operational in scope. From a methodological point, we will first test the hypothesis that kriging can be adapted to the complexity of the data involved in BUVARD, according to Chen et al. (2013) and Roustant et al. (2018). Kriging on mixed variables will be compared to regression (which has already shown efficiency in water quality modeling), additive modeling (which is based on non parametric statistics), and kriging on quantitative variables only (one metamodel per qualitative level). The aim is not to provide a comprehensive review of metamodel performances but to evaluate the ability of

kriging to take into account a mix of qualitative and quantitative variables and to provide high prediction performances. From an operational perspective through several applications in the context of risk analysis and management, we will then show that metamodeling is an interesting tool for addressing current operational issues of optimal vegetative filter strip design.

2. Material and methods

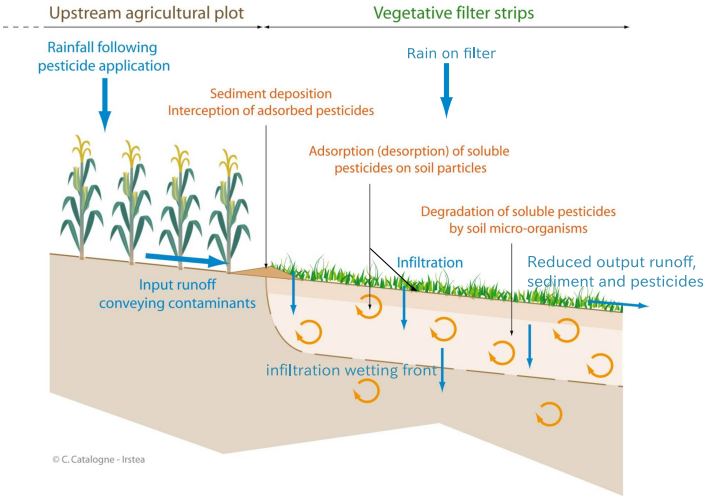
2.1. Modeling toolkit description

Vegetative filter strips, when properly designed and implemented, reduce surface runoff from upslope fields by improving soil infiltration, and thus slowing down pesticide transfer due to water runoff from the plot to the downslope water body (see Figure 1a). While they are also effective for sediment and pesticide trapping, this study focuses on hydrological processes, assuming that surface runoff is the main process driving pesticide transfer in a VFS (Carluer et al., 2017). BUVARD is a set of coupled models dedicated to designing VFSs by simulating their efficiency to limit runoff transfer (Carluer et al., 2017). Based on the benchmark numerical model VFSMOD (Muñoz-Carpena et al., 1999; Muñoz-Carpena and Parsons, 2004; Muñoz Carpena et al., 2018; Lauvernet and Muñoz-Carpena, 2018) that quantifies dynamic effects of VFS site-specific pesticide mitigation efficiency, BUVARD offers a full framework for designing VFSs over France. It is based on representative rainfall and surface runoff events for a specific site, given the local climate and the contributing area characteristics (slope, length, land use, humidity state - these four parameters determine the curve number, a parameter reflecting soil tendency to generate runoff). The design method is then based on (i) the quantification of water flows produced by the contributing area with the Curve Number method (an empirical parameter describing the potential surface runoff generation of the contributing area, USDA-NRCS, 1986), and (ii) VFS capacities to infiltrate incoming flows, running VFSMOD (see Figure 1b). Finally, simulations are run for several lengths of the VFS, and the optimal one is selected according to the required efficiency level (for example, 70% runoff reduction). In total, the user has to enter more than 70 input factors to run the BUVARD toolkit.

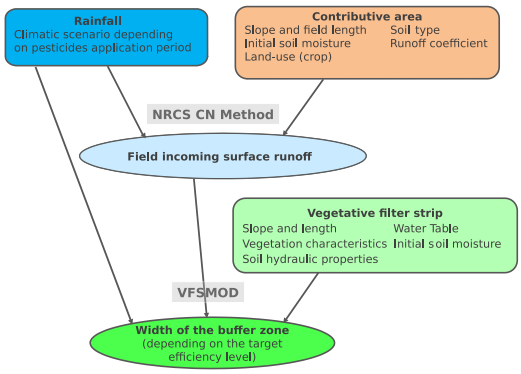
To simplify the design procedure in practice, we defined some typical soil types to describe hydrodynamic inputs instead of hydrodynamic parameter values, including saturated hydraulic conductivity and soil retention curve parameters required by VFSMOD. The four main classes of the four VFS soil types implemented in BUVARD correspond to clay loam (CLO), sandy clay loam (SCL), silt loam (SIL) and sandy loam (SAL), according to Brown et al. (2012). However, these soil type descriptions are based on the assumption that VFSs have been established for years, since filter maturity has an effect on the soil structure at the surface (linked to root development and biological activity) and the resulting increase in infiltration capacity. This influence on newly implanted VFSs is taken into account in BUVARD for predominantly clay soils (clay loam -labeled 'clo'- and sandy-clay-loam -labeled 'scl'). It is considered less significant on other types of soils, which already have relatively high infiltration properties. 'clo' and 'scl' soils have the same retention curve parameters as,

respectively, 'CLO' and 'SCL', but higher permeability and porosity. Their characteristics are given in Table 1 and described in detail in Carluer et al. (2017).

The local climate, if not available to the user, can also be represented by rainfall types included in BUVARD, which are built from a classification method of rainfall events with 1-year return period. A return period of 1 year means that these rainfall events are likely to occur once a year on average. Therefore, sizing will be at least effective for smaller and common rains, representing the majority of the precipitation that falls each year. However, it will be less accurate for more exceptional events (for which buffer zone devices are not generally designed). The rainfall classification differentiates the temporal structures of rainfall events according to their duration, season and geographical location, based on the analysis of small time step observations made by four rainfall stations considered as representative of four major climatic zones in metropolitan France and Corsica (see Figure 2). For each region this method generates some so-called "rainfall event types" summarized in Table 2: one short summer episode (D = 1 h, labeled S01), one long summer episode (D = 6 h, labeled S06), one short winter episode (D = 2 h, labeled W02), one long winter episode (D = 12 h, labeled W12). More details on the rainfall classification method over France can be found in Appendix 8.1 and in Catalogne et al. (2016a,b). These two categorical variables (soil type and rainfall event type) were implemented in BUVARD to simplify its use and the inputs settings and have some consequences on the metamodeling method to be developed, such as the need to adapt kriging correlation functions to categorical variables in lieu of the usual quantitative variables.



(a) Main processes occurring on a vegetative filter strip.



(b) Sizing method flowchart of the BUVARD modeling toolkit.

Figure 1: Processes occurring on a vegetative filter strip and simulated in the BUVARD tool based on VFSMOD.

Level	soil type	K_{sat} (cm/day)	K_{sat} (m/s)	θ_{sat}	θ_r	n_{VG}	α_{VG} (m^{-1})
clo	clay loam	6.72	7.78E-07	0.428	0.083	1.45	1.01
scl	sandy clay loam	12.89	1.49E-06	0.398	0.065	1.36	1.91
SIL	silt loam	23	2.66E-06	0.458	0.068	1.66	0.54
CLO	clay loam	49.49	5.73E-06	0.456	0.083	1.45	1.01
SCL	sandy clay loam	53.93	6.24E-06	0.495	0.065	1.36	1.91
SAL	sandy loam	98.63	1.14E-05	0.478	0.055	1.44	2.4

Table 1: Soil type characteristics required by VFSMOD for the shallow water table condition. n_{VG} and α_{VG} are soil retention curve parameters (Van Genuchten). This is assumed here that m_{VG} from Van Genuchten equation is defined as $m_{VG} = 1-1/n_{VG}$. θ_{sat} and θ_r are respectively saturated and residual soil saturation. Soil type written in uppercase relates to an established filter that was implemented for more than 5 years, and lowercase relates to newly implanted filters.

Level	type of event	duration
S01	short summer episode	1 h
S06	long summer episode	6 h
W02	short winter episode	2 h
W12	long winter episode	12 h

Table 2: Rainfall types description. h stands for hour(s).

2.2. Metamodeling methodology

In what follows, the numerical code is represented by a function f such that:

$$\begin{array}{rccc} f: & [0,1]^p & \longrightarrow & [a;b] \subset \mathbb{R} \\ & (x_1, ..., x_p) & \longmapsto & f(x_1, ..., x_p) \end{array} \tag{1}$$

where x_1, ..., x_p denote the input variables, also called input parameters in some applications.

In the context of computer experiments, Gaussian process modeling, also called kriging, is widely used (Santner et al., 2003). This metamodel has the advantage of being interpolant, which is well-suited for numerical code deterministic outputs. It is also of interest since it evaluates a prediction uncertainty as well as the prediction itself. In this work we propose comparing kriging to two other metamodeling approaches, linear and generalized additive models, in the specific statistical context of a mix of quantitative and qualitative variables. The first is traditional linear regression which is a robust parametric learning method. The regression functions are given in advance and are generally first and second degrees of input variables and their interactions. For the generalized additive model (GAM), the regression functions are estimated by a non-parametric approach, such as splines. This overview is not complete but a good starting point to show the relevance of kriging with mixed variables and its ability to adapt to this particular application.

2.2.1. Kriging-based metamodeling with mixed variables

Gaussian process modeling

In the context of Gaussian process modeling the function f is assumed to be a realization of a Gaussian process $(Y(\mathbf{x}))_{(\mathbf{x})}$ with a constant mean m and a stationary covariance function k that can be written $\forall \mathbf{x}, \mathbf{x}', k(\mathbf{x}, \mathbf{x}') = cov(Y(\mathbf{x}), Y(\mathbf{x}')) = \sigma^2 r(\mathbf{x} - \mathbf{x}')$, where r is a correlation function and σ^2 the variance of the field (Santner et al., 2003). In the literature, r is often considered to be a tensor product of one-dimensional parametric correlation functions.

Parameter estimation

Let $(\mathbf{x}^1, \ldots, \mathbf{x}^n)$ be the initial design of experiments, where $\mathbf{x}^k \in [0,1]^p$. Let $\mathbf{y} = (y^1, ..., y^n)$ be the evaluation of the numerical code at points $(\mathbf{x}^1, \ldots, \mathbf{x}^n)$. The correlation parameters, the variance σ^2, and the mean m are estimated by maximum likelihood from the observations.

Kriging predictions

Let R be the correlation matrix between observation points $Y(\mathbf{x}^1)$ ...$Y(\mathbf{x}^n)$, i.e. $R[k,l] = r(\mathbf{x}^k - \mathbf{x}^l)$. Let $\mathbf{x}$ be a new point where the function f has to be predicted. Let $r_{\mathbf{x}}$ be the vector composed of the correlations between $Y(\mathbf{x})$ and $Y(\mathbf{x}^1)$ ...$Y(\mathbf{x}^n)$.

The kriging mean and kriging variance are the conditional expectation and variance, and are given by the following equations:

$$\hat{Y}(\mathbf{x}) = \mathbb{E}\left(Y(\mathbf{x})|Y(\mathbf{x}^1) = y^1, ..., Y(\mathbf{x}^n) = y^n\right) = m + r_{\mathbf{x}}^t R^{-1}\left(\mathbf{y} - m\mathbf{1}\right)$$

$$\hat{\sigma}^2(\mathbf{x}) = \mathbb{V}\left(Y(\mathbf{x})|Y(\mathbf{x}^1) = y^1, ..., Y(\mathbf{x}^n) = y^n\right) = \sigma^2\left(1 - r_{\mathbf{x}}^t R^{-1} r_{\mathbf{x}}\right)$$

The prediction formulas indicate that:

- The prediction mean is a weighted average of the observations. The weight of each observation y_i depends on the correlation between $Y(\mathbf{x})$ and $Y(\mathbf{x}^i)$, i.e. on the distance between the observed point $\mathbf{x}^i$ and the point $\mathbf{x}$ to be predicted.
- The prediction variance is zero at observation points and increases as the distance to the observation points increases.

Correlation functions adapted to the mixing of qualitative and quantitative variables

When inputs are a mixture of nominal, ordinal, and quantitative inputs, the usual tensor product correlation function can be written:

$$r((\mathbf{x}, \mathbf{w}, \mathbf{z}) - (\mathbf{x}', \mathbf{w}', \mathbf{z}')) = r_{quanti}(\mathbf{x} - \mathbf{x}')r_{ordi}(\mathbf{w} - \mathbf{w}')r_{quali}(\mathbf{z} - \mathbf{z}')$$

where $\mathbf{x} = (x_1, ..., x_{p_\mathbf{x}}) \in [0,1]^{p_\mathbf{x}}$ represents quantitative variables, $\mathbf{w} = (w_1, ..., w_{p_\mathbf{w}}) \in \Xi_1 \times ... \times \Xi_{p_\mathbf{w}}$ represents ordinal variables, and $\mathbf{z} = (z_1, ..., z_{p_z}) \in \Xi_1 \times ... \times \Xi_{p_\mathbf{z}}$ represents nominal variables.

For quantitative inputs, there are many correlation kernels, from non differentiable to infinity differentiable (Santner et al., 2003). In what follows we consider the two differentiable Matérn 5/2 kernel that depends on one parameter θ_j in each direction:

$$r_{quanti}(\mathbf{x} - \mathbf{x}') = \prod_{j=1}^{p_\mathbf{x}} \left(1 + \frac{\sqrt{5}|x_j - x_j'|}{\theta_j} + \frac{5(x_j - x_j')^2}{3\theta_j^2}\right) \exp\left(-\frac{\sqrt{5}|x_j - x_j'|}{\theta_j}\right)$$

Sometimes an underlying order exists between levels. In this case, the variable is an ordinal variable instead of a categorical variable. Its correlation structure is a classical kernel for continuous variables composed with a non-decreasing transformation F (the cumulative distribution function of the standard Gaussian distribution). In the case of a Matérn kernel, the formula is as follows, depending on a unique parameter ν_j in each direction:

$$r_{ordi}(\mathbf{w} - \mathbf{w}') = \prod_{j=1}^{p_\mathbf{w}} \left(1 + \frac{\sqrt{5}|F(w_j) - F(w_j')|}{\nu_j} + \frac{5(F(w_j) - F(w_j'))^2}{3\nu_j^2}\right) \exp\left(-\frac{\sqrt{5}|F(w_j) - F(w_j')|}{\nu_j}\right) \quad (2)$$

with $w_j, w_j' \in \Xi_j$ the set of ordered levels.

Correlation kernels for nominal inputs are less commonly used. Based on Chen et al. (2013), three different correlation functions have been compared:

1. **cov-quali-isotropic**:

$$r(z_j - z'_j) = exp\left(-\rho \mathbf{1}_{z_j \neq z'_j}\right) \tag{3}$$

2. **cov-quali-product**:

$$r(z_j - z'_j) = exp\left(-(\rho_{z_j} + \rho_{z'_j})\mathbf{1}_{z_j \neq z'_j}\right) \tag{4}$$

3. **cov-quali-anisotropic**:

$$r(z_j - z'_j) = exp\left(-\rho_{z_j, z'_j}\mathbf{1}_{z_j \neq z'_j}\right) \tag{5}$$

The global formula is then as follows:

$$r_{quali}(\mathbf{z} - \mathbf{z}') = \prod_{j=1}^{p_{\mathbf{z}}} r(z_j - z'_j) \tag{6}$$

with $z_j, z'_j \in \Xi_j$ the set of categorical levels and r picked in the above list. Note that the number of parameters depends on the correlation choice.

As an example, Table 3 shows how the three covariance structures can be represented for one nominal input described by 4 different levels.

Kernel	4 levels	Number of correlation parameters
cov-quali-isotropic	$\begin{pmatrix} 1 & \rho & \rho & \rho \\ & 1 & \rho & \rho \\ & & 1 & \rho \\ & & & 1 \end{pmatrix}$	1
cov-quali-product	$\begin{pmatrix} 1 & \rho_1\rho_2 & \rho_1\rho_3 & \rho_1\rho_4 \\ & 1 & \rho_2\rho_3 & \rho_2\rho_4 \\ & & 1 & \rho_3\rho_4 \\ & & & 1 \end{pmatrix}$	4
cov-quali-anisotropic	$\begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ & 1 & \rho_{23} & \rho_{24} \\ & & 1 & \rho_{34} \\ & & & 1 \end{pmatrix}$	6

Table 3: Kernels used for kriging with nominal input: example of an input described by 4 levels

.

2.2.2. Other metamodels

In this study, Gaussian process modeling is compared to linear model and GAM, as these latter models have been often used in water quality modeling.

Linear models

In linear models, the most appropriate model is a model with interactions and quadratic terms. The general formula is as follows:

$$y^k = \beta_0 + \Sigma_{j=1}^{p} \beta_j x_j^k + \Sigma_{j,j'} \beta_{jj'} x_j^k x_{j'}^k + \epsilon^k \tag{7}$$

where $\epsilon^1, ..., \epsilon^n$ are i.i.d. random variables with a centered Gaussian distribution of variance σ^2. In the presence of qualitative inputs, the linear model is developed in equation 10 provided in the Appendix.

As the formula is composed of a large number of predictors, there is a risk of overfitting. To avoid this problem, complexity is reduced by selecting a sub-model. This is obtained by minimizing the BIC criterion (Schwarz, 1978), expressed as follows:

$$BIC = -2\log(\mathcal{L}(y^1, \ldots, y^n)) + \log(n)K$$

where K is the number of predictors and $\mathcal{L}$ the likelihood of the observations. BIC carries out a trade-off between data fitting and parsimony.

In the case of qualitative inputs (nominal and/or ordinal), the formulas remain the same, each level is represented by an indicator function. Identifiability is then possible thanks to identifiability constraints (Fabozzi et al., 2014).

Generalized additive models (GAM)

GAM (Wood, 2017) are interesting visualization tools that allow the user to understand the dependencies between the inputs and the output.

We propose a brief review of how GAM works. With the most simple formula, in the context of GAM, f is approximated by function $\tilde{f}$ such that:

$$\tilde{f}(\mathbf{x}) = \alpha + g_1(x_1) + \ldots + g_p(x_p). \tag{8}$$

$\tilde{f}$ is then the sum of different functions built in each direction of the input space. α represents the global mean of the function, while all the functions g_j are assumed orthogonal and of null integral. The model is based on a non-parametric function estimation. A backfitting algorithm provides an efficient estimation of f (Friedman et al., 2008). A more complex formula would involve functions of interactions between inputs. In high dimension, the problem is to add the most efficient inputs or combinations of inputs to prevent excessive

increase in model complexity.

In the case of qualitative inputs (nominal and/or ordinal) the estimation of non-parametric functions does not make sense. These inputs are introduced in GAM in the same way as they are in the linear model.

2.2.3. Design of experiments

The chosen design of experiments is an n-point Latin hypercube sampling (LHS). A LHS is obtained by dividing a hypercube in some n equal intervals in each direction: one sampled point is chosen per interval such that its projection in any direction is unique per level. For the dimensions of the hypercube wich refer to the qualitative variables, the sampling is made in integer space, and the sampled integers are associated to the levels of the qualitative variables. By doing this, we ensure that the quantitative variables and qualitative variables are sampled in the same LHS. Latin Hypercube Sampling is not optimal in terms of space filling when the dimension is higher than 1. Another dispersion property must be optimized in higher dimension. In this case, the chosen design is a LHS optimized according to maximin criterion Φ_{Mm}, which definition is as follows: let $X_n = \{\mathbf{x}^1, ..., \mathbf{x}^n\}$ be a design of experiments, $\Phi_{Mm}(X_n) = min_{i \neq j} \|\mathbf{x}^i - \mathbf{x}^j\|$.

Φ_{Mm} maximizes the minimal distance between each pair of points from the sample, using the Simulated Annealing routine with a pure geometric algorithm (Pronzato and Muller, 2012; Dupuy et al., 2015).

2.2.4. Quality criterion

To quantify the metamodel quality, two data sets are generated: a n points training set (also called design of experiments) and a N points test set. The performances of the different models (linear, GAM, and kriging) can be compared through different measures (Dupuy et al., 2015). In this paper, Q2 is chosen because of its simplicity of interpretation as a percentage of explained variance (the closer to 1, the better):

$$\text{Q2} = 1 - \frac{\sum_{k=1}^{N} (y^k - \hat{y}^k)^2}{\sum_{k=1}^{N} (y^k - \overline{y})^2} \tag{9}$$

where $\overline{y}$ is the observation mean and $\hat{y}^k$ is the prediction of the kth point of the test set. The prediction is obtained either by the linear model, GAM, or kriging. In the hydrological literature, this quantity was introduced as model efficiency NSE (Nash and Sutcliffe, 1970). Ritter and Muñoz Carpena (2013) analyzed the NSE and introduced the FITEVAL statistical hypothesis testing procedure for model acceptability (NSE$>$ 0.65) and model "pedigrees" (unacceptable, acceptable, good, very good) based on NSE values.

3. Metamodeling of the BUVARD tool

3.1. Metamodeling setup

3.1.1. Metamodel output and inputs

The metamodel will be built on an area based on the experimental catchment Yzeron (Lagouy et al., 2015), which is included in the climatic zone 2 (see Figure 2). The target output is not the optimal VFS size but the runoff delivery ratio, RDR, an index representing the efficiency of the VFS in reducing surface runoff, relative to rainfall volume, and that is directly a VFSMOD output (Muñoz-Carpena and Parsons, 2004):

$$\text{RDR} = \frac{\text{runoff exiting the filter}}{\text{runoff entering the filter+rainfall}}$$

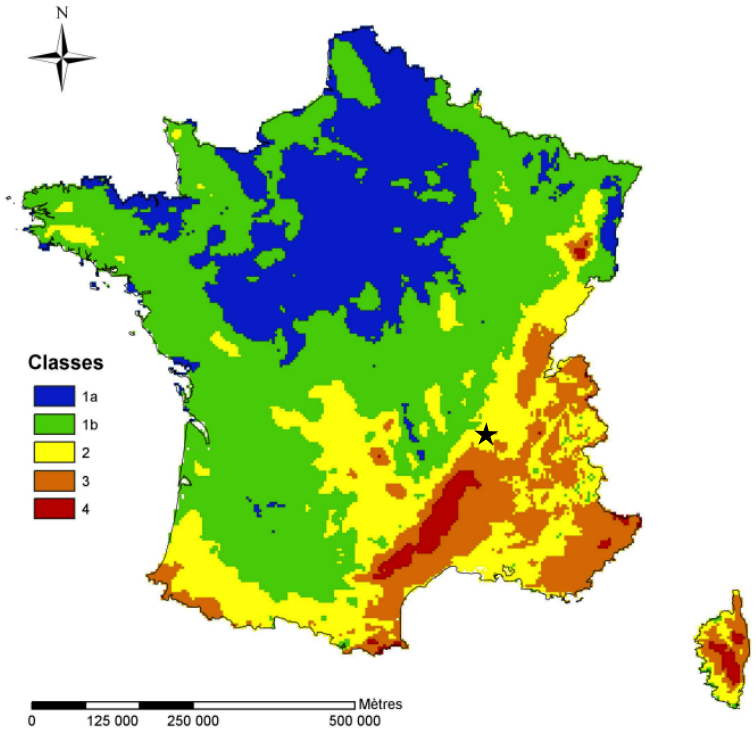


Figure 2: Mapping of the five climate classes obtained following the upward hierarchical classification from Catalogne et al. (2016b). The black star indicates the case study region, close to Lyon (Yzeron catchment).

The very last step in BUVARD consists in getting the "optimal" VFS size according to the selected efficiency. This is achieved by running the metamodel on several VFS sizes, computing the efficiency $(1 - \text{RDR})$, and selecting the closest one to the objectives of the study. This allows flexibility on the chosen efficiency, since the user will be able to set it once the metamodel has been built.

Sampling is performed on 7 inputs: 5 quantitative and 2 qualitative, with ranges listed in Table 4. The 5 quantitative inputs are the curve number (CN), the slope of the contributing area (Slope), the length of the

contributing area, the water table depth (WTD) below the filter, and the length (VL) of the filter. Note that filter *length* may be called filter *width* in some studies: it is the dimension of the filter in the 1D surface runoff direction. The two qualitative inputs are the rainfall typical event (rainfall type) with 4 levels, and the filter soil type (VFS soil) with 6 levels. This represents 24 distinct combinations.

Quantitative input	interval	unit	element
Curve number (CN)	[63, 99]	-	CA
Slope	[0.1, 20]	-	CA
Length	[25, 300]	[m]	CA
Water table depth (WTD)	[0.50, 4]	[m]	VFS
Filter length (VL)	[3, 30]	[m]	VFS
Qualititative input	**levels**		**element**
Rainfall type	S01 - S06 - W02 - W12		CA and VFS
VFS soil	clo - scl - SIL - CLO - SCL - SAL		VFS

Table 4: List of inputs. The boundaries (respectively list of levels) are given for quantitative (respectively qualitative) inputs. CA stands for contributive area, and VFS for vegetative filter strip.

3.1.2. Training and test samples

The design of experiments is a maximin LHS composed of 100 points in the quantitative hypercube space per couple of (qualitative) levels. Since there are 24 different pairs of levels (VFS soil type $\times$ rainfall type), the total number of points in the training sample is 24×100. The test sample is based on an independent LHS, built on 40 points, leading to a size of 24×40. The effect of the sampling size is explored in Section 3.2.3.

3.1.3. Metamodel formulas

A reference kriging model is considered by building separately one independent kriging for each couple of qualitative inputs. While this method is the easiest, it can be time-consuming to implement, depending on the number of qualitative inputs and their levels. In this study, there are 6 levels for the soil types and 4 for the rainfall types, which implies building separately 24 metamodels, each based on a sampling of 100 parameter settings. In addition, this approach does not take any advantage of information available from other levels, which can be an important limitation when the number of simulations is restricted.

Three other kriging models are built, taking into account a mixture of quantitative and qualitative inputs. They are based on the correlation kernels introduced in Section 2.2.1 and are adapted to the inputs listed in Table 4. The formulas are described in Appendix 8.2.

These models are implemented in the R software with package *kergp*[4].

The estimated linear model is composed of all the linear terms and the interactions between the inputs (CN, Slope, length, WDT, VL, VFS soil, and rainfall type) and of the quadratic terms for the quantitative inputs (CN, Slope, length, WDT, VL). The formula is introduced in Appendix 8.3. An automatic selection of relevant terms by BIC criterion is then applied to this complete formula.

The GAM formula for the case study involves principal effects but also major interactions between qualitative and quantitative inputs. The formula is detailed in Appendix 8.4.

3.2. Evaluation of the methods

3.2.1. Comparison of kriging kernels

Table 5 gives the Q2 (NSE) values obtained for the four kriging-based approaches listed above. Building 24 independent kriging models is the worst method with a Q2 (NSE) = 0.939, which is already very high and proves that kriging is efficient. Specific correlation structures improve kriging quality, especially **cov-quali-product**. **Cov-quali-isotropic** would also be interesting to evaluate further when the training size is smaller, since it needs less parameters to estimate. However, **cov-quali-product** is retained for the comparison with the linear model and GAM.

Kriging correlation kernels	Q2 (NSE)
by group of levels	0.9391
cov-quali-isotropic	0.9636
cov-quali-product	0.9641
cov-quali-anisotropic	0.9416

Table 5: Q2 (NSE) on the test sample according to different correlation functions.

3.2.2. Performance of metamodeling methods accounting for rainfall and soil levels

In this section, we compare the kriging model selected in the previous section to a linear model and GAM. The formula used for the linear model and the procedure for selecting only significant terms is explained in Section 2.2.2.

Predicted values are plotted against true values in Figure 3. The densities of predicted errors on the test sample are compared in Figure 4. Only errors greater than 5% are considered. It can be observed that kriging does better than linear and additive models. Q2 (NSE) on the test set with kriging is around 96%, whereas it is

[4] `https://CRAN.R-project.org/package=kergp`

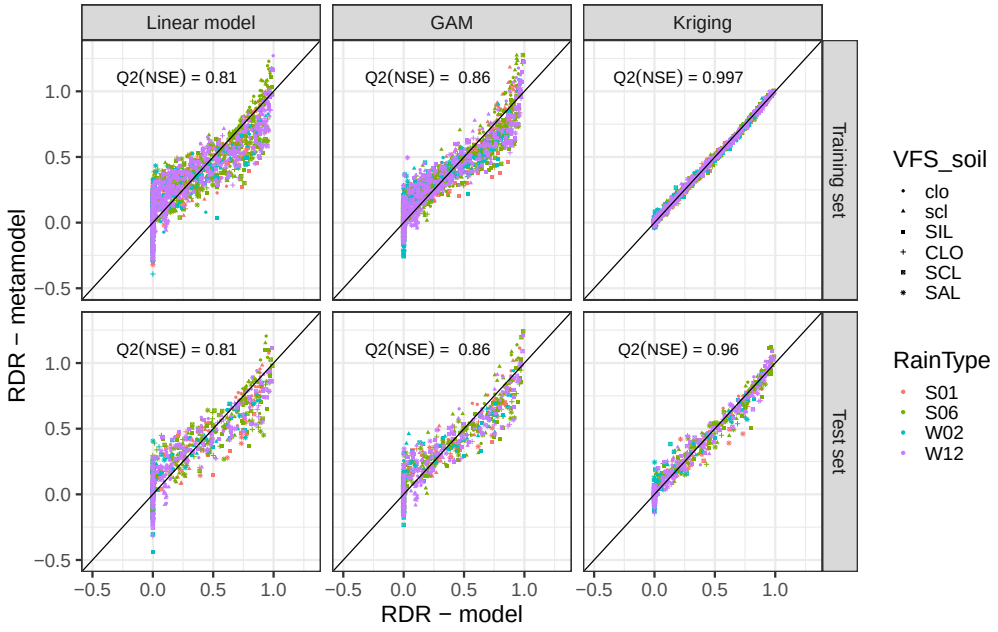


Figure 3: Comparison between RDR from the model and from the metamodel built on the training set, accounting for qualitative inputs (6 VFS soils and 4 rainfall types) with 3 different methods: linear model, generalized additive model (GAM), and kriging (with a **cov-quali-product** covariance function). The predicted RDR vs the observed RDR is compared on the training set (top panels) and on the test set (bottom panels).

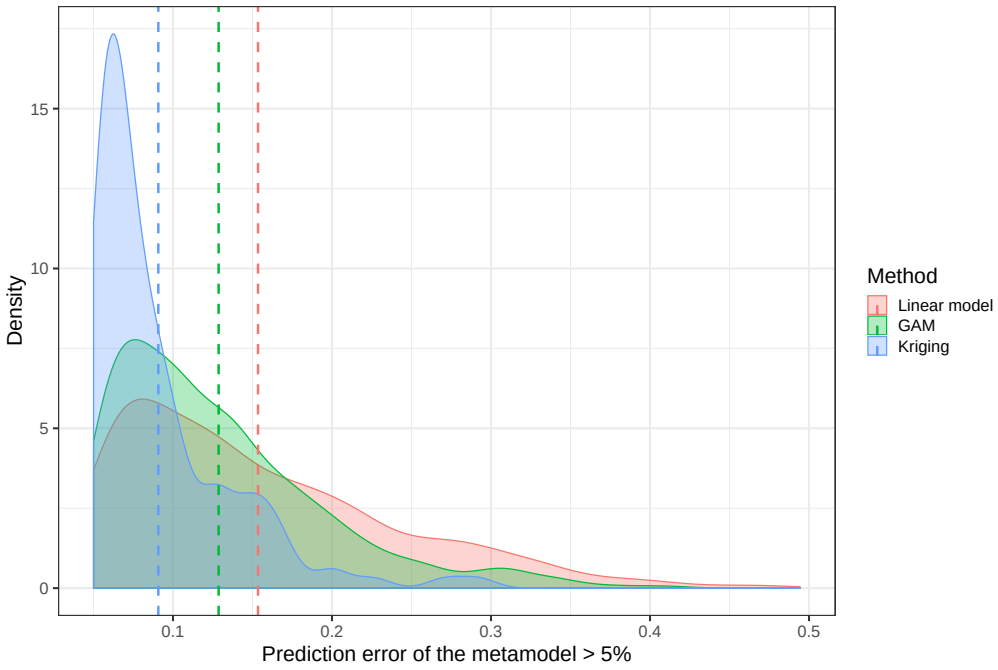


Figure 4: Densities of the errors (larger than 5%) of the metamodel computed for the test set for the 3 different methods: linear model, GAM and kriging. Errors are defined as the difference between RDR simulated by the model and RDR predicted by the metamodel. The dotted lines represent the mean error for the corresponding metamodel.

only equal to 86% for the additive model and only slightly over 80% for the linear model. Furthermore, errors are much more concentrated next to zero. Average prediction errors are 0.15 for the linear model, 0.13 for GAM , and 0.09 for Gaussian process modeling. Kriging is a semi-parametric model, very smooth, as opposed to the linear model. It fits data very well without hard assumptions. Note that kriging is an interpolation method. Q2 (NSE) should exactly be equal to one on the training set. This is not the case here since we add a small error variance on the diagonal of the covariance matrix as explained in Roustant et al. (2012). This is known to smooth the response surface and to ease likelihood optimization and parameter estimation.

Finally, the FITEVAL statistical hypothesis testing procedure was applied for all the metamodels (Ritter and Muñoz Carpena, 2013). As expected, all methods past the statistical test for acceptable model (NSE>0.65) (Table 6). Also, as expected, the pedigree of the metamodels (or the Model quality) is different. This supports the choice of kriging that shows a "very good" model quality.

Method	NSE	p-value	RMSE	Model quality
Linear	0.811 [0.771 - 0.846]	0.000	0.143 [0.135 - 0.153]	ACCEPTABLE to GOOD
GAM	0.863 [0.838 - 0.882]	0.000	0.122 [0.115 - 0.129]	GOOD
Kriging	0.964 [0.955 - 0.971]	0.000	0.063 [0.058 - 0.069]	VERY GOOD

Table 6: Evaluation of metamodels following the FITEVAL method. NSE and RMSE (median value and 95% confidence interval in brackets) represent the metamodel goodness-of-fit with the original model.

3.2.3. Effect of sampling size

The effect of training sampling size on the metamodeling quality of the different metamodels was tested on kriging with mixed inputs, and compared to the linear model and GAM (Figure 5). The sampling size parameter was tested from 10 to 100 per pair of levels and should be multiplied by 24 pairs of levels (4 rainfall types * 6 soil types) to obtain the total sampling size. The criterion Q2 (NSE) is performed on an independent test dataset of size 40*24. We observe that the comparison of metamodels is stable with the changes in size and with a clear advantage for kriging. Indeed, the averaged Q2 (NSE) is 0.78 for the linear model, 0.85 for GAM, and 0.93 for kriging. For the smallest sampling size (10*24), the formula for building GAM had to be optimized with fewer terms than for other sizes. For 10*24 points, kriging and GAM are close and improve quality compared to the linear model. It can also be observed in Figure 5 that sensitivity to the training sample size is weak above 30 points per couple of levels. However, if the trend is examined closely, the explained variance increases monotonically for kriging (2 points of variance gain from 0.94 with 50 points to 0.96 with 100 points), whereas it is not that clear for the two other models. This is accounted for by the structure of the metamodels. The kriging predictor is a local average of observations in the vicinity of the point. The more points are added, the more relevant the predictor is. The linear model and GAM depend

on formulas built from a relevant terms selection. All parameters and the most important pairs of interaction are included, but some terms are most likely missing and it is not sufficient to achieve the explained kriging variance level even when sampling size increases.

Finally, kriging by level (or modality) is another way of evaluating the relevance of mixed inputs in the kriging method. Whatever the sampling size, the kriging model combining qualitative and quantitative inputs gives much higher Q2s (NSEs). For very small designs, the quality of kriging by level is unstable and decreases sharply.

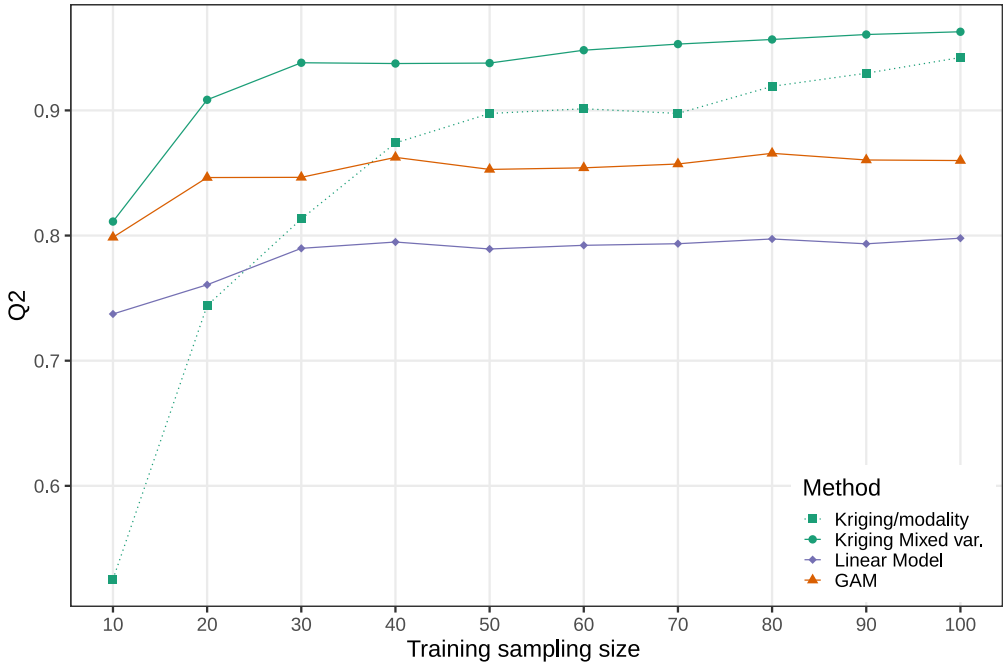


Figure 5: Effect of training sampling size on metamodel quality with 3 different methods: linear model, GAM, and kriging (mixed and per level). The evaluation criterion is Q2 (NSE) on the test sample. For each case, the total number of points in the sample is the size defined in the x-label multiplied by 24 (pairs of levels).

4. Applications for risk analysis and management

4.1. Using kriging as a tool for uncertainty quantification

In this section, the analyses are obtained on the kriging model adjusted from the 100-point maximin LHS. Figure 6 shows a sensitivity study of the runoff delivery ratio to the 5 quantitative inputs (CN, Slope, Length, WTD, VL). Each boxplot corresponds to the distribution of RDR according to the variation of one quantitative input on its definition range, while the four others are kept fixed to their mean value defined in Table 4.

CN is particularly noticeable from this analysis, generally generating a much wider output distribution than other inputs. It can therefore be assumed that the uncertainty associated with CN, which is a rather

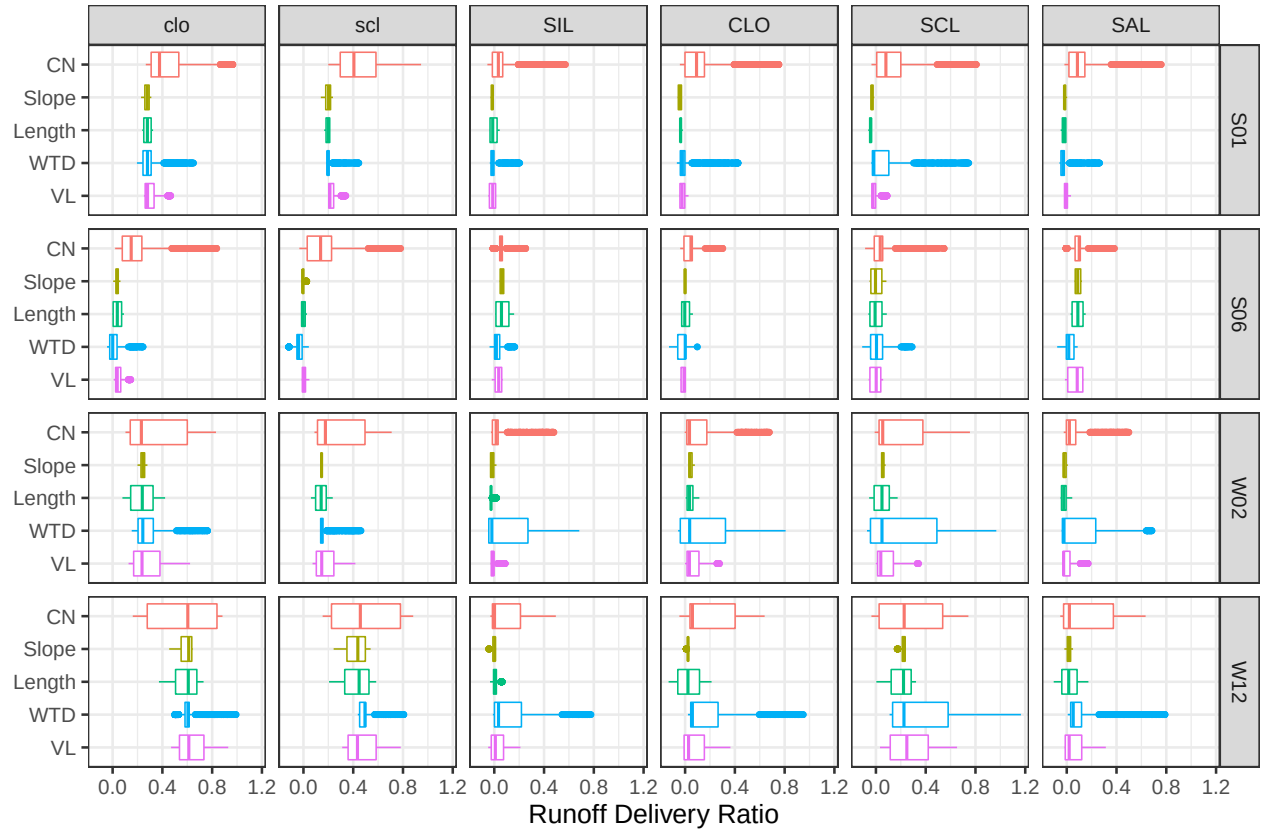


Figure 6: Effect of quantitative input uncertainty on the RDR, depending on the qualitative input levels (6 VFS soils and 4 rainfall types - see Tables 1 and 2). The plotted distributions are obtained by predicting kriging estimated from the 100-point maximin LHS (the total number of training points is $24 * 100$) with the cov-quali-isotropic covariance function.

difficult parameter to estimate, propagates significantly over the estimated efficiency of the VFS. The shallow water table depth (WTD) and the length of the VFS (VL) play a significant role in uncertainty, depending on the soil and rain combinations (especially in case of winter rains, and newly implanted VFS soils for the length of the VFS). We continue to observe a large number of outliers with distributions related to CN. This implies that low CN combined with low rainfall events can cause zero VFS efficiency (RDR close to 1) quite roughly but also quite rarely. Finally, slope and length of the contributing area have a similar behavior, spreading little uncertainty and thus having little influence on RDR.

These OAT sensitivity analysis results should be interpreted with caution and are not a substitute for a global sensitivity analysis based on variance decomposition, for example. However, they are quite consistent with what is known about the global sensitivity analysis results assessed in previous studies with the VFSMOD model in particular (Lauvernet and Muñoz-Carpena, 2018; Muñoz-Carpena et al., 2007, 2010). Finally, they confirm that CN must be handled with care, precisely when it is difficult to define, and that the results used to design a grass strip from a simulation of BUVARD with a single CN value must always be accompanied by an uncertainty associated with this value.

4.2. Using the metamodel as a tool for testing regulations

In France as in many countries, the same VFS length is recommended whatever the context, i.e., whatever the climate, the soil type, the contributing area length, etc. (see e.g., [5]). The metamodel, due to its ease of use, can help to evaluate the accuracy of this one-fits-all rule in an operational context. For example, one can wonder what would be the probability of failure if the rule stipulated a 10 m long filter in all cases. In this case, failure is defined by obtaining a filter efficiency of less than 70% of surface runoff reduction. Figure 7 shows that the probability of failure with the 10 m long rule greatly depends on the types of rain and soil. It is particularly risky for short summer (1 h) and long winter (12 h) events and for newly implanted VFSs with clay loam and sandy clay soils. On the contrary, with established VFSs with clay loam and sandy clay soils, the length of 10 m has low probability of getting an RDR upper than 30%, although still dependent on rainfall events. Globally, the risk of failing is more than 25% with this decision.

The same test was performed with a 5 m long filter and showed that we take a higher but similar risk, depending on the same types of soil and climate, than for a 10 m long filter. Metamodeling is thus a useful tool for discussing environmental policy with stakeholders and decision makers, and for testing this kind of one-fits-all rule.

[5] Order of 27 December 2019 - , amending the Order of 4 May 2017 on the placing on the market and use of plant protection products and their adjuvants referred to in Article L. 253-1 of the Rural and Maritime Fisheries Code

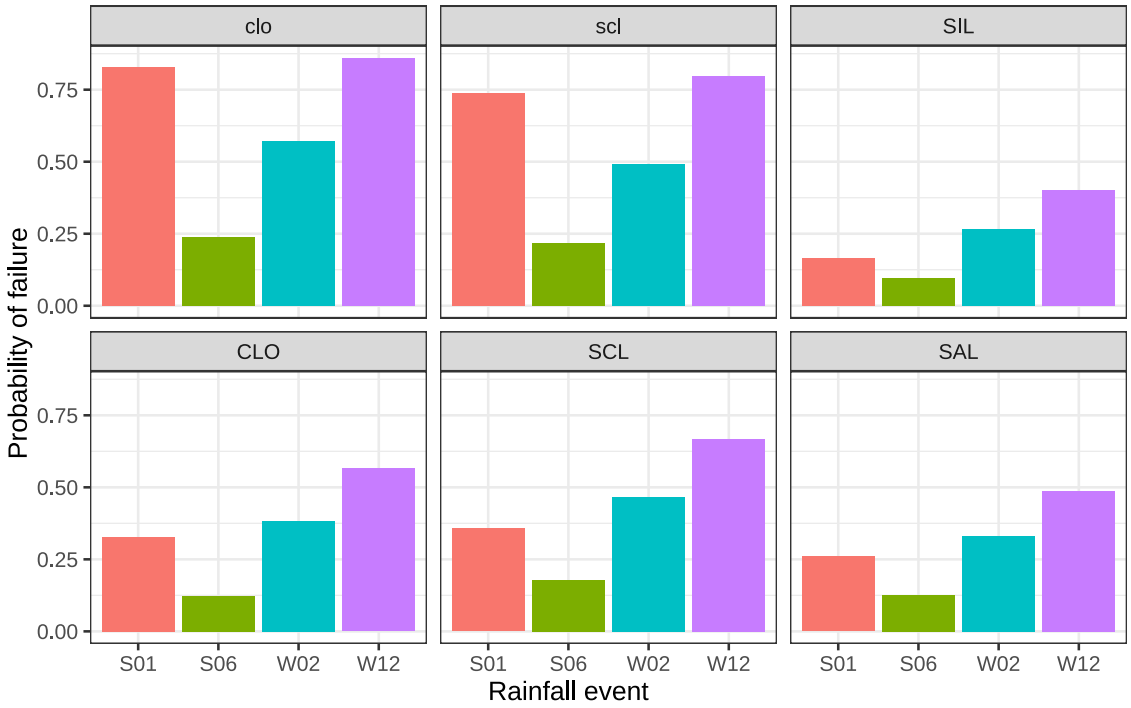


Figure 7: Probability of failure by applying the 10 m long rule for any vegetative filter strip, for different soil and rainfall types.

4.3. Using the metamodel as a tool for local operational purposes

Finally, we propose focusing on how to use the tool correctly for the case where the user knows the site well, accounting for uncertainties related to the site or the (un)available information. Let us consider that the user knows the average slope of the contributing area (here Slope = 10%), the length of the contributing area (here Length = 160 m), the average water table depth in winter (here WTD = 0.8 m) and in summer (here WTD = 2.5 m). We chose two contrasting types of soil combined with an event, based on the probability of failure when choosing the 10 m rule (Figure 7): an established VFS with sandy loam soil (SAL) with the event Summer 6 h, *i.e.*, a case where the probability of failure was less than 12%, and a newly established VFS with clay loam soil (clo) with the event Winter 12h, *i.e.*, where the probability of failing was close to 80%. Considering that the curve number is a very influential parameter on VFS efficiency but also a very uncertain one, we let it vary in relevant ranges considering the soil type ($[63, 75]$ for the sandy loam soil, and $[70, 80]$ for the clay loam soil), and tried to find the optimal length VL considering all these input data. Figure 8 shows that, whatever the value of VL and CN, filter efficiency is much lower (RDR is larger) in winter on the clo soil than in summer on the SAL soil. If we consider the worst case (CN at its highest value), the length of the VFS must exceed 30 m to reach a reasonable RDR (lower than 0.7) for the clo soil. On average, the sandy soil shows a much higher efficiency whatever the filter length and CN due to its high permeability, as opposed to the clay soil, which is never more efficient than 50% and even generally at 20% to 30% of efficiency.

These graphical representations are an example of the potential use of this tool as a demonstration support for training, consulting, or technical analysis. It provides a clear way of addressing the uncertainty associated with the VFS sizing, in a domain that is not yet familiar with it.

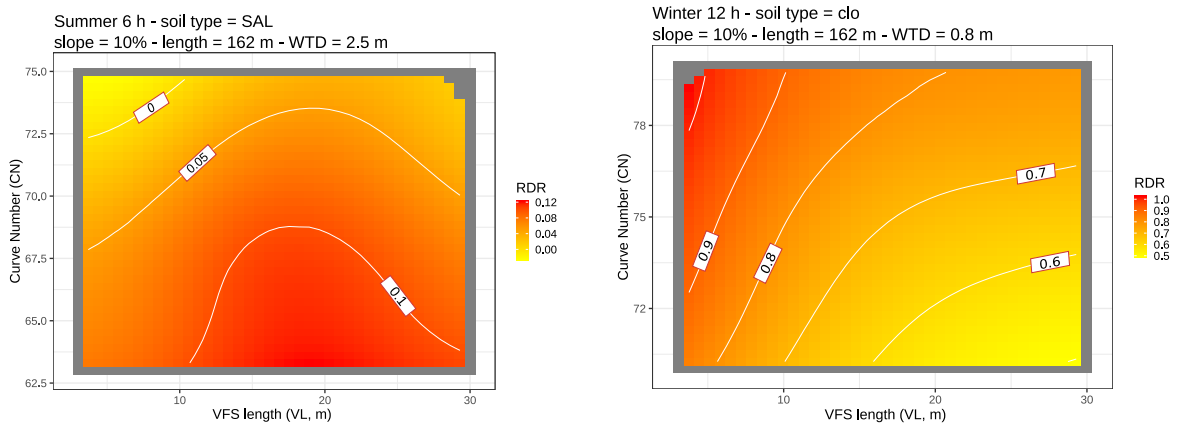


Figure 8: RDR predicted by the metamodel considering fixed slope, length of the field, and water table depth. Left panel: VFS soil type is sandy loam soil (SAL), CN is sampled in the range $[63 - 75]$. Right panel: VFS soil type is newly implanted clay loam soil (clo), CN is sampled in the range $[70 - 80]$.

5. Conclusions

In this paper we are concerned with designing vegetative filter strips to reduce river water pollution as a consequence of pesticide runoff from upslope fields. Although VFS characteristics can be computed directly with the available BUVARD framework, in practice this modeling tool is difficult to handle for users. We introduce model reduction techniques to reduce BUVARD complexity and help users design VFSs in a simplified metamodeling context. Relevant information is summed up into a few quantitative and qualitative inputs. The first objective of this work consists in finding the best metamodeling techniques to represent the numerical model. The results show that kriging with a correlation kernel adapted to a mixture of qualitative and quantitative inputs outperforms the linear model and GAM. The second objective is to show that a surrogate model can help users design VFSs for a specific site (soil type and rainfall event) and can also account for uncertainty. The study showed that the metamodel is able to perform simulations close to the original BUVARD ones to help size vegetative filter strips, and that it can perfom a large number of simulations at low cost. Whereas simulation costs with the BUVARD tool are low but dependent on the specific problem (e.g., shallow water table and high rainfall), the metamodel predictions have a fixed cost regardless of the conditions, cost which is nearly negligible.

A first step of uncertainty quantification is proposed here. The surrogate model is a powerful tool for gaining system information. Future research should focus on global sensitivity analysis to quantify the relative

influence of each input and their interactions. Secondly, the metamodel can be used to optimize filter length in different situations taking into account uncertainty. The idea is to propose a procedure that automatically and sequentially adds BUVARD simulations to find the best VFS under uncertainties. Several studies are available and could be adapted to our context. For example, in Janusevskis and Le Riche (2013), uncertainty is summarized through an expectation. Our future work will consist in quantifying uncertainty by a global sensitivity analysis and resolving the problem with a robust optimization procedure. For example, this could be a multi-objective optimization of both expectation and variance of the output.

Another important research topic in metamodeling concerns numerical codes which involve functional inputs or outputs. In BUVARD, the problem was avoided by summarizing the dynamic phenomena using variables that integrate the dynamics, for instance via rainfall event types. However, physically, the system evolves with time, and most environmental models contain true dynamics, as is the case in VFSMOD, the model on which BUVARD is based. It would be interesting to consider and to model the intermediate temporal outputs. Other methods could also be explored considering this issue, such as polynomial chaos or random forest. Finally, even though the processes occurring within a vegetative filter strip are interactively complex in space and time, the BUVARD metamodel was able to account for these processes to achieve a good predictive quality. This study showed that reducing complex models with appropriate methods offers the opportunity to apply research tools in risk pollution assessment or decision-making.

6. Acknowledgments

The authors would like to thank Esperan Panodou for his help on the mixgp package, Robert Faivre and Nicolas Forquet for fruitful discussions on metamodeling issues, Rafael Muñoz-Carpena for his invaluable advice on how to improve the latest version of the paper, Clotaire Catalogne, Nadia Carluer, and Guy Le Hénaff for discussions about the BUVARD tool and VFS sizing and Benjamin Renard for the stimulating debates on non-linearities. This work is part of the BUVARD_MM project which was supported by a CNRS-INSU LEFE-MANU grant.

7. References

Adriaanse, P. I., Leerdam, R. C. V., Boesten, J. J., 2017. The effect of the runoff size on the pesticide concentration in runoff water and in FOCUS streams simulated by PRZM and TOXSWA. Science of the Total Environment 584–585, 268 – 281.

Arnaud, P., Lavabre, J., 2002. Coupled rainfall model and discharge model for flood frequency estimation. Water Resources Research 38 (6), 11–1–11–11.
URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2001WR000474

Brown, C., Balderacchi, M., Van Beinum, W., Capri, E., Trevisan, M., 2012. Definition of vegetative filter strip scenarios for Europe. Tech. rep., Environment Department, University of York, Heslington, York, YO10 5DD, UK, p. 71.

Carluer, N., Lauvernet, C., Noll, D., Muñoz-Carpena, R., 2017. Defining context-specific scenarios to design vegetated buffer zones that limit pesticide transfer via surface runoff. Science of the Total Environment 575, 701 – 712.

Catalogne, C., Lauvernet, C., Carluer, N., 2016a. Some abacus to size vegetative buffer strips to limit pesticide transfer by runoff - from a research tool to final users. In: 46e congrès du Groupe Français des Pesticides, 17/05/2016-19/05/2016, Bordeaux, FRA. p. 2.
URL http://cemadoc.irstea.fr/cemoa/PUB00053253

Catalogne, C., Lauvernet, C., Leblois, E., 2016b. Analyse des structures temporelles de pluies pour la définition de hyetogrammes en entrée de la chaine de dimensionnement des bandes tampons végétalisées BUVARD. Tech. rep., Irstea.
URL http://irsteadoc.irstea.fr/cemoa/PUB00055314

Chen, X., Wang, K., Yang, F., 2013. Stochastic kriging with qualitative factors. Vol. Proceedings of the 2013 Winter Simulation Conference. IEEE, pp. 790–801.

Dosskey, M. G., Helmers, M. J., Eisenhauer, D. E., 2006. An approach for using soil surveys to guide the placement of water quality buffers. Journal of Soil and Water Conservation 61 (6), 344–354.
URL http://www.jswconline.org/content/61/6/344

Dosskey, M. G., Helmers, M. J., Eisenhauer, D. E., 2011. A design aid for sizing filter strips using buffer area ratio. Journal of Soil and Water Conservation 66 (1), 29–39.
URL http://www.jswconline.org/content/66/1/29

Dubois, A., Parisse, S., 2017. Pesticides : évolution des ventes, des usages et de la présence dans les cours d'eau depuis 2009 auteur(s): Editeur: Sdes date de publication: 2017-03 numéro de la publication: 94 issn:. URL https://www.statistiques.developpement-durable.gouv.fr/pesticides-evolution-des-ventes-des-usag

Dupuy, D., Helbert, C., Franco, J., 2015. DiceDesign and DiceEval: two R packages for design and analysis of computer experiments. Journal of Statistical Software, Articles 65 (11), 1–38. URL https://www.jstatsoft.org/v065/i11

Fabozzi, F., Focardi, S., Rachev, S., Arshanapalli, B. G., 2014. The Basics of Financial Econometrics: Tools, Concepts, and Asset Management Applications. John Wiley and Sons, Ch. Regression Models with Categorical Variables.

Fienen, M. N., Nolan, B. T., Feinstein, D. T., Starn, J. J., 2015. Metamodels to bridge the gap between modeling and decision support. Groundwater 53 (4), 511–512.

Fox, G., Muñoz-Carpena, R., Purvis, R., 2018. Controlled laboratory experiments and modeling of vegetative filter strips with shallow water tables. Journal of Hydrology 556 (Supplement C), 1 – 9.

Friedman, J., Hastie, T., Tibshirani, R., 2008. The Elements of Statistical Learning : Data Mining, Inference and Prediction. Springer Series in Statistics. Springer-Verlag, New York.

Fu, B., Merritt, W. S., Croke, B. F., Weber, T. R., Jakeman, A. J., 2019. A review of catchment-scale water quality and erosion models and a synthesis of future prospects. Environmental Modelling and Software 114, 75 – 97.

Gatel, L., Lauvernet, C., Carluer, N., Weill, S., Paniconi, C., Dec 2019a. Sobol global sensitivity analysis of a coupled surface/subsurface water flow and reactive solute transfer model on a real hillslope. Water 12 (1), 121. URL http://dx.doi.org/10.3390/w12010121

Gatel, L., Lauvernet, C., Carluer, N., Weill, S., Tournebize, J., Paniconi, C., 2019b. Global evaluation and sensitivity analysis of a physically based flow and reactive transport model on a laboratory experiment. Environmental Modelling & Software 113, 73 – 83. URL http://www.sciencedirect.com/science/article/pii/S1364815217310204

Higdon, D., 2015. Discussion of "Computer experiments with qualitative and quantitative variables: A review and reexamination". Quality Engineering 27 (1), 14–16.

Janusevskis, J., Le Riche, R., 2013. Simultaneous kriging-based estimation and optimization of mean response. Journal of Global Optimization 55 (2), 313.

Lacas, J.-G., Voltz, M., Gouy, V., Carluer, N., Gril, J.-J., 2005. Using grassed strips to limit pesticide transfer to surface water: a review. Agronomy for Sustainable Development 25 (2), 253–266.

Lagouy, M., Branger, F.and Thollet, F., Breil, P., Dramais, G., 2015. Suivi hydrologique du bassin versant périurbain de l'yzeron; irstea.
URL https://www.data.gouv.fr/fr/datasets/suivi-hydrologique-du-bassin-versant-periurbain-de-lyzeron

Lauvernet, C., Muñoz-Carpena, R., 2018. Shallow water table effects on water, sediment, and pesticide transport in vegetative filter strips – part 2: model coupling, application, factor importance, and uncertainty. Hydrology and Earth System Sciences 22 (1), 71–87.

Muñoz Carpena, R., Lauvernet, C., Carluer, N., 2018. Shallow water table effects on water, sediment, and pesticide transport in vegetative filter strips – part 1: nonuniform infiltration and soil water redistribution. Hydrology and Earth System Sciences 22 (1), 53–70.

Muñoz-Carpena, R., Fox, G., Sabbagh, G., 2010. Parameter importance and uncertainty in predicting runoff pesticide reduction with filter strips. Journal of Environmental Quality 39 (2), 630–41.

Muñoz-Carpena, R., Parsons, J. E., 2004. A design procedure for vegetative filter strips using VFSMOD-W. Transactions of the ASAE 47 (6), 1933–1941.

Muñoz-Carpena, R., Parsons, J. E., Gilliam, J. W., 1999. Modeling hydrology and sediment transport in vegetative filter strips. Journal of Hydrology 214 (1-4), 111–129.

Muñoz-Carpena, R., Zajac, Z., Kuo, Y.-M., 2007. Global sensitivity and uncertainty analyses of the water quality model VFSMOD-W. Transactions of the ASABE 50 (5), 1719–1732.

Nash, J., Sutcliffe, J., 1970. River flow forecasting through conceptual models part I - A discussion of principles. Journal of Hydrology 10 (3), 282–290.
URL http://www.sciencedirect.com/science/article/B6V6C-487FF7C-1XH/2/4bba4b13a68e4942542254b2ec23796b

Piñeros Garcet, J., Ordoñez, A., Roosen, J., Vanclooster, M., 2006. Metamodelling: Theory, concepts and application to nitrate leaching modelling. Ecological Modelling 193 (3), 629 – 644.

Pronzato, L., Muller, W., 2012. Design of computer experiments: space filling and beyond. Statistics and Computing 22, 681–701.

Rasmussen, C. E., 2006. Gaussian processes for machine learning. MIT Press.

Ratto, M., Castelletti, A., Pagano, A., 2012. Emulation techniques for the reduction and sensitivity analysis of complex environmental models. Emulation techniques for the reduction and sensitivity analysis of complex environmental models 34, 1–4.

Reichenberger, S., Bach, M., Skitschak, A., Frede, H.-G., 2007. Mitigation strategies to reduce pesticide inputs into ground- and surface water and their effectiveness; A review. Science of The Total Environment 384 (1–3), 1–35.

Ritter, A., Muñoz Carpena, R., 2013. Performance evaluation of hydrological models: statistical significance for reducing subjectivity in goodness-of-fit assessments. Journal of Hydrology 480 (1), 33–45.

Roustant, O., Ginsbourger, D., Deville, Y., 2012. DiceKriging, DiceOptim: two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization. Journal of Statistical Software, Articles 51 (1), 1–55.
URL https://www.jstatsoft.org/v051/i01

Roustant, O., Padonou, E., Deville, Y., Clement, A., Perrin, G., Giorla, J., Wynn, H., 2018. Group kernels for gaussian process metamodels with categorical inputs. Accepted in SIAM/ASA Journal on Uncertainty Quantification.
URL https://hal.archives-ouvertes.fr/hal-01702607/file/groupKernels.pdf

Rouzies, E., Lauvernet, C., Barachet, C., Morel, T., Branger, F., Braud, I., Carluer, N., 2019. From agricultural catchment to management scenarios: A modular tool to assess effects of landscape features on water and pesticide behavior. Science of The Total Environment 671, 1144 – 1160.
URL http://www.sciencedirect.com/science/article/pii/S0048969719310381

Santner, T. J., Williams, B. J., Notz, W. I., 2003. The design and analysis of computer experiments. Springer Series in Statistics. Springer-Verlag, New York.
URL http://dx.doi.org/10.1007/978-1-4757-3799-8

Schwarz, G., 03 1978. Estimating the dimension of a model. Ann. Statist. 6 (2), 461–464.
URL https://doi.org/10.1214/aos/1176344136

Shirmohammadi, A., Chaubey, I., Harmel, R. D., Bosch, D. D., Muñoz-carpena, R., Dharmasri, C., Sexton, A., Arabi, M., Wolfe, M. L., Frankenberger, J., Graff, C., Sohrabi, T. M., 2006. Uncertainty in tmdl models. Trans. ASABE, 49–1033.

Tiktak, A., Boesten, J., Van Der Linden, A., Vanclooster, M., 2006. Mapping ground water vulnerability to pesticide leaching with a process-based metamodel of EuroPEARL. Journal of Environmental Quality 35 (4), 1213–1226, cited By 28.

USDA-NRCS, 1986. Urban hydrology for small watersheds. Natural Resources Conservation Service, Conservation Engineering Division, Technical Release 55 (TR-55) (Second ed.).

Villa-Vialaneix, N., Follador, M., Ratto, M., Leip, A., 2012. A comparison of eight metamodeling techniques for the simulation of N_2O fluxes and N leaching from corn crops. Environmental Modelling and Software 34, 51–66.

White, M. J., Arnold, J. G., 2009. Development of a simplistic vegetative filter strip model for sediment and nutrient retention at the field scale. Hydrological Processes 23 (11), 1602–1616.

Wood, S. N., 2017. Generalized Additive Models: An Introduction with R, 2nd edition. Texts in Statistical Science. Chapman & Hall CRC Press.

Younes, A., Voltz, M., Fahs, M., Guellouz, L., Mara, T., Musa Baalousha, H., 2018. A new efficient Bayesian parameter inference strategy: Application to flow and pesticide transport through unsaturated porous media. Journal of Hydrology 563, 887–899.

Zhang, Y., Notz, W. I., 2015. Computer experiments with qualitative and quantitative variables: A review and reexamination. Quality Engineering 27 (1), 2–13.

8. Appendices

8.1. Rainfall event definition

This appendix provides more details on the rainfall event acquisition over France for the BUVARD method. Large climatic zones have been defined in France, considered to be "homogeneous" and for which typical temporal rainfall structures are computed on the basis of rainfall stations deemed representative of each zone (see Figure 2) (Catalogne et al., 2016b). These areas were defined using a hierarchical bottom-up classification (Ward method), the input values of which are the intensity values of the four types of rainfall episode previously defined over France. The values are estimated by the SHYREG method from the seasonal intensity-duration-frequency curves and characterized by an intensity I, a duration D (1, 2, 3, 3, 4, 6, 12, 24, 48, 72 h), and a seasonalized 1-year return period (i.e. a value for the winter period from December to May and a value for the summer period from June to November) (Arnaud and Lavabre, 2002). Table 7 gives the quantiles on the Summer and Winter durations selected in this study on the four climatic zones. The analysis protocol then consists in extracting all the episodes of duration D and intensity I for the season in question (summer or winter) on real data from each zone, from fine-time step rainfall chronicles (cumulative over 5 minutes).

Climatic class	Site	S01	S06	W02	W12
1a	Orgeval (77)	10.1	21.1	8.7	18.9
1b	Bourville (76)	11.7	24.7	8.7	23.1
2	Yzeron (69)	18.1	35.7	11.9	32.6
3-4	Roujan (34)	23	46.7	18.2	48.4

Table 7: Rainfall quantiles (in mm) on the 4 climatic classes on 4 events of one year return-period: S is for summer events and W for winter events. 01 to 12 is the rainfall event duration in hours.

8.2. Kriging kernels adapted to BUVARD metamodeling setup

The three kernels which are compared in the study are as follows, where $x, x' \in [0,1]^5$ represent two vectors of the quantitative inputs (CN, Slope, Length, WDT, VL), $w, w' \in \{clo, scl, SIL, CLO, SCL, SAL\}$ for VFS Soil and $z, z' \in \{S01, S06, W02, W12\}$ for rainfall type:

- cov-quali-isotropic: the global kernel is a product of **Matérn kernel** for quantitative inputs and **cov-quali-isotropic** for VFS soil type and rainfall type (equation 3).

$$r((x,w,z)-(x',w',z')) = r_{quanti}(x-x')r_{quali1}(w-w')r_{quali1}(z-z')$$

  This kernel depends on 7 range parameters θ: 5 for the quantitative part, 1 for VFS soil and 1 for rainfall type.

- cov-quali-product: the global kernel is a product of **Matérn kernel** for quantitative inputs and **cov-quali-product** for VFS soil type and rainfall type (equation 4).

$$r((x,w,z)-(x',w',z')) = r_{quanti}(x-x')r_{quali2}(w-w')r_{quali2}(z-z')$$

  This kernel depends on 15 range parameters θ: 5 for the quantitative part, 6 for VFS soil, and 4 for rainfall type.

- cov-quali-anisotropic: the global kernel is a product of **Matérn kernel** for quantitative inputs and **cov-quali-anisotropic** for rainfall type (equation 5) and **cov-ordinal** for VFS soil type (equation 2).

$$r((x,w,z)-(x',w',z')) = r_{quanti}(x-x')r_{ordi}(w-w')r_{quali3}(z-z')$$

  This kernel depends on 15 range parameters θ: 5 for the quantitative part, 1 for VFS soil, and 6 for rainfall type. The order for soils was determined by increasing permeability:
  clo, scl, SIL, CLO, SCL, SAL.

8.3. Linear model with qualitative inputs

In linear models, the most appropriate model is a model with interactions and quadratic terms. The general formula is as follows:

$$\begin{aligned} y^k &= \nu + \Sigma_{i=1}^{p} \beta_i x_i^k + \Sigma_{i \leq i'} \beta_{ii'} x_i^k x_{i'}^k \\ &+ \Sigma_{j=1}^{q} \Sigma_{h=1}^{l_j} \alpha_{jh} \mathbf{1}_{z_j^k = s_{jh}} + \Sigma_{j \leq j'} \Sigma_{h \leq h'} \alpha_{jhj'h'} \mathbf{1}_{z_j^k = s_{jh}} \mathbf{1}_{z_{j'}^k = s_{j'h'}} \\ &+ \Sigma_{i=1}^{p} \Sigma_{j=1}^{q} \Sigma_{h=1}^{l_j} \gamma_{ijh} x_i^k \mathbf{1}_{z_j^k = s_{jh}} + \epsilon^k \end{aligned} \tag{10}$$

where $\epsilon^1, ..., \epsilon^n$ are i.i.d. variables with a centered Gaussian distribution of variance σ^2.

8.4. GAM formula with qualitative variables

The following terms have been selected:

- splines in the directions: CN, slope, length, WTD, and VL,
- linear additivity in rainfall type and VFS soil type,
- splines in the interactions between rainfall types and CN, WTD, and length,
- splines in the interaction between VFS soil type and CN,
- linear interaction between VFS soil type and rainfall type,
- splines in interactions VL*CN, CN*length, and VL*length.

The R formula has the following form (see Wood, 2017, for an R implementation):
gam_fun = gam(RDR $\sim s(CN)+s(Slope)+s(Length)+s(WTD)+s(VL)+RainType+VFS\_soil+s(CN, by = RainType) + s(WTD, by = RainType) + s(Length, by = RainType) + s(CN, by = VFS\_soil) + RainType : VFS\_soil + s(VL, CN) + s(CN, Length) + s(VL, Length), data = training, method = "ML")$