

---

# Using input warping to improve the Bayesian optimisation of a complex epidemiological model of the sharka virus

Victor Picheny · Coralie Picard · Gaël Thébaud

Received: date / Accepted: date

**Abstract** Optimizing black-box numerical models remains a challenge in many research fields. In this work, we focus on a Bayesian optimization approach, accounting for local invariances of the model with respect to its input variables. More precisely, we incorporate the prior knowledge that the model is insensitive to variations of some of its input variables when other input variables take a particular value. To this end, we propose a new warping technique applied to the parameter space that encode the invariances. This approach is tested on a simulation model of sharka disease spread and management that exhibits several invariances. We analyze the contribution of the warping on the Bayesian optimization of sharka control options. We show that the warping step significantly improves the rate of convergence of the BO algorithm.

**Keywords** Bayesian optimisation, warping, spatio temporal model, sharka

## 1 Introduction

Mathematical models are increasingly used in many research fields to understand and optimize a process. For instance, they are useful in epidemiology to predict epidemics and to propose efficient control options [4, 5, 18, 33, 1, 14, 34, 9]. However, these epidemiological studies are mostly focused on improving one control option which generally depends on only one or two parameters in their model, although various control actions are usually applied simultaneously to manage an epidemic. All these actions could be jointly optimized but taking into account numerous management

---

V. Picheny  
MIAT, Université de Toulouse, INRA, Castanet-Tolosan, France  
Tel.: +33561285551  
E-mail: victor.picheny@inra.fr

C. Picard  
BGPI, Montpellier SupAgro, INRA, Univ. Montpellier, Cirad, TA A-54/K, 34398, Montpellier

G. Thébaud  
to do

parameters in an optimization problem can be difficult, especially when the management efficiency depends on the interaction between these parameters.

In this study, we analyse a simulation model of sharka disease spread and management. This disease, caused by a virus transmitted by aphids, is one of the most damaging diseases of stone fruit trees belonging to the genus *Prunus* (e.g. peach, apricot and plum) [3, 25]. Our model includes epidemiological parameters which vary between simulations, and various landscapes on which the virus can spread, which means that this model is stochastic. In addition, management parameters allow to simulate orchard surveillance. Here, we aim to optimize these management parameters using a efficient optimization algorithm.

Within the wide range of potential approaches to solve such optimization problems, black-box optimization methods have proven to be popular in this context [28], in particular because they are in essence non-intrusive: they only require pointwise evaluations of the model at hand (output value for a given set of inputs), as opposed to knowing the underlying mechanisms of the model, structural information, derivatives, etc. This greatly facilitates implementation and avoids developing tailored algorithms. In this work, we focus more particularly on the so-called *Bayesian optimization* (BO) approaches [17, 30], which are well-suited to tackle stochastic and expensive models.

In some cases, the user possesses relevant information regarding his model that could facilitate the optimization task. Accounting for this information within a black-box optimization framework (or rather: *grey box*) may be a challenging task as it is, in essence, unnatural. In this work, we focus on a particular type of information, which we refer to as *local invariance*: for some values of a subset of parameters, it is known that the model is insensitive to another subset of parameters. As an illustration, take a function  $y$  that depends on two discs, parameterized by  $r_1 \in [0, r_{\max}]$  (radius of the first disc) and  $r_2 \in [0, r_{\max}]$  (radius of the second disc) with  $r_1 > r_2$ . An action  $A_1$  is conveyed on the first disc and another action  $A_2$  on the second. Setting  $r_1 = 0$ , we have  $r_2 = 0$ , thus for any value of  $A_2$ ,  $y$  is not impacted. Taking into account such invariances would avoid wasting computational resources exploring those regions. Moreover, it would avoid the problem of having local *plateaus* of the optimization landscape, that are likely to slow down the optimization process or even prevent convergence to an optimum.

Intuitively, one may want to rework the definition of the parameters to optimize over in order to remove the invariances. However (as we show in 2), such a reformulation is not always possible. Here, we propose to keep the optimisation problem unchanged, and convey the invariance information to the BO algorithm directly, by applying a *warping* [31, 32] to the parameter space.

The remainder of this paper is structured as follow. Section 2 describes the sharka model and its invariances. Section 3 presents the basics of Bayesian optimization and our warping strategy. Finally, section 4 analyses the efficiency of the warping on the sharka model.

## 2 Model description and problem set-up

The simulation model that we analyze in this work is a stochastic, spatially explicit, SEIR (susceptible-exposed-infectious-removed) model that simulates sharka spread and management actions [including surveillance, removals and replantations 22, 26, 27].

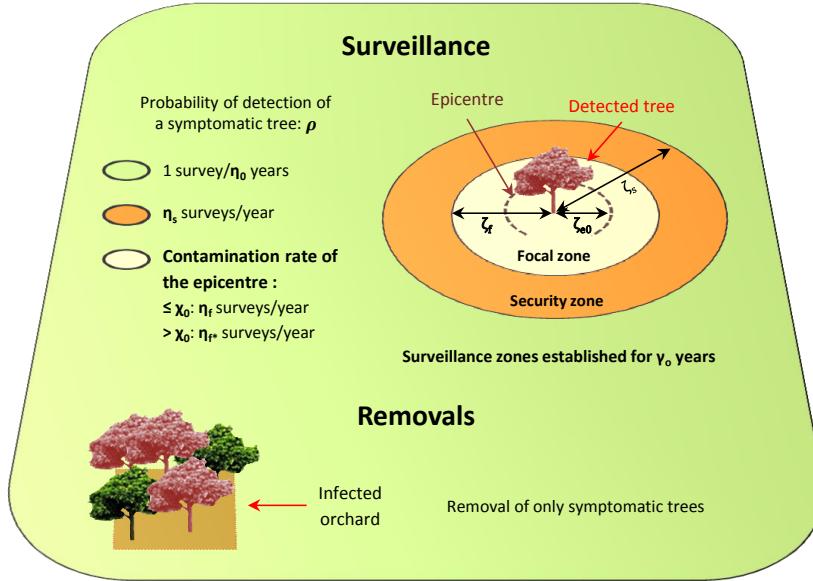
This model is orchard-based, with a discrete time step of one week. It allows to perform simulations on landscapes composed of uncultivated areas and patches on which peach trees are grown. The patches can be more or less aggregated in the landscape however, we only use in this work the 30 landscapes with a high level of patch aggregation as described by Picard et al. [19]. During the simulation, the trees in the patches are characterized by different states. When the simulation begins, they are not infected: they are in the “susceptible” state. Then, the virus is introduced the first year of the simulation in one of the patches and spreads through orchards (new introductions can also occur during the entire simulation on all patches). The virus causes changes in tree status: from “susceptible”, they become “exposed” (infected but not yet infectious or symptomatic), “infectious hidden” (after the end of the latent period), “infectious detected” (when specific symptoms are detected on the tree during a survey), and “removed” (when the tree is removed from the patch). The model output is an economic criterion, the net present value (NPV), which accounts for the benefit generated by the cultivation of productive trees and the costs induced by fruit production and disease management [27].

In order to simulate wide range of epidemic and management scenarios, the model includes 6 epidemiological and 23 management parameters [27, 19]. In this work, we will use the 6 epidemiological parameters and only 10 management parameters to performed some optimizations quickly. Among the 23 management parameters, we removed parameters corresponding to plantation restrictions, removals, and surveillance of young orchards. The parameters we kept include distances of 3 zones for which the surveys are more or less frequent as well as their duration, the probability of the infected tree detection, and a contamination threshold which can request to increase the surveillance frequency in a focal zone. Details of management parameters used in this study are presented in Fig.1 and Table 1 (this table also includes the variation ranges of the parameters in the model).

Here, we aim to optimize the management strategy of the disease (i.e. to find the combination of management parameters allowing to obtain the best NPV), taking into account the epidemic stochasticity. However, we note that some combinations of management parameters can represent the same management, which may cause problems in the optimization process. Indeed, we observe that some management parameters are not useful when other parameters have a value of 0, which means that they can take any values without modifying the simulation. For example, when a zone radius is 0, the associated surveillance frequency have no impact on the NPV (regardless its value). The methodological developments that are proposed in this work address this issue by removing the parameter combinations which lead to the same management. The parameter invariances removed from the model are listed in Table 2.

**Table 1** Management parameters implemented in the previously developed model with minimum and maximum values corresponding to the variation range of each parameter.

		Min	Max
$\rho$	Probability of detection of a symptomatic tree	0	0.66
$\gamma_o$	Duration of observation zones (year)	0	10
$\zeta_s$	Radius of security zones (m)	0	5800
$\zeta_f$	Radius of focal zones (m)	0	5800
$\zeta_{eo}$	Radius of observation epicenter (m)	0	5800
$1/\eta_0$	Maximal period between 2 observations (year)	1	15
$\eta_s$	Observation frequency in security zones ( $\text{year}^{-1}$ )	0	8
$\eta_f$	Observation frequency in focal zones ( $\text{year}^{-1}$ )	0	8
$\eta_{f*}$	Modified observation frequency in focal zones ( $\text{year}^{-1}$ )	0	8
$\chi_o$	Contamination threshold in the observation epicenter, above which the observation frequency in focal zone is modified	0	1



**Fig. 1** Management actions implemented in the model.

### 3 Methods: Bayesian optimization

#### 3.1 Overview

Bayesian optimization can be seen as a modernization of the statistical response surface methodology for sequential design [2], where the basic idea is to replace an

**Table 2** Invariances of management parameters. For instance, when  $\gamma_O = 0$  or when  $\rho = 0$ ,  $\chi_o$  does not influence the model output.

Management parameters	OR	OR	OR
$\chi_o$	$\gamma_O = 0$	$\rho = 0$	
$\zeta_{eo}$	$\gamma_O = 0$	$\zeta_s = 0$	$\rho = 0$
$\zeta_f$	$\gamma_O = 0$	$\zeta_s = 0$	
$\eta_{f*}$	$\gamma_O = 0$	$\rho = 0$	
$\zeta_s$	$\gamma_O = 0$	$\eta_s = 0$	
$\eta_s$	$\gamma_O = 0$		
$\eta_f$	$\gamma_O = 0$		

expensive-to-evaluate function by a cheap-to-evaluate surrogate one. In BO, Gaussian process (GP) regression, or kriging, is used to provide flexible response surface fits. GPs are attractive in particular for their tractability, since they are simply characterized by their mean  $m(\cdot)$  and covariance (or kernel)  $k(\cdot, \cdot)$  functions, see e.g., Rasmussen and Williams [24]. In the following, we consider zero-mean processes ( $m = 0$ ) for the sake of conciseness.

Conditionally on  $n$  noisy observations  $\mathbf{f} = (f_1, \dots, f_n)$ , with independent, centered, Gaussian noise, that is,  $f_i = y(\mathbf{x}_i) + \varepsilon_i$  with  $\varepsilon_i \sim \mathcal{N}(0, \tau_i^2)$ , the predictive distribution of  $y$  is another GP, with mean and covariance functions given by:

$$\mu(\mathbf{x}) = \mathbf{k}(\mathbf{x})^\top \mathbf{K}^{-1} \mathbf{f}, \quad (1)$$

$$\sigma^2(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') - \mathbf{k}(\mathbf{x})^\top \mathbf{K}^{-1} \mathbf{k}(\mathbf{x}'), \quad (2)$$

where  $T$  denotes the tranposition operator,  $\mathbf{k}(\mathbf{x}) := (k(\mathbf{x}, \mathbf{x}_1), \dots, k(\mathbf{x}, \mathbf{x}_n))^\top$  and  $\mathbf{K} := (k(\mathbf{x}_i, \mathbf{x}_j) + \tau_i^2 \delta_{i=j})_{1 \leq i, j \leq n}$ ,  $\delta$  standing for the Kronecker function.

Commonly,  $k(\cdot, \cdot)$  belongs to a parametric family of covariance functions such as the Gaussian and Matérn kernels, based on hypotheses about the smoothness of  $y$ . Corresponding hyperparameters are often obtained as maximum likelihood estimates, see e.g., Rasmussen and Williams [24] or Roustant et al [29] for the corresponding details.

BO typically tackles optimization problems of the form:

$$\begin{aligned} \min \quad & y(\mathbf{x}) \\ \text{s.t. } & \mathbf{x} \in \mathbb{X}, \end{aligned}$$

with  $\mathbb{X} \in \mathbb{R}^d$  is usually a bounded hyperrectangle and  $y : \mathbb{R}^d \rightarrow \mathbb{R}$  is a scalar-valued objective function.

Optimization amounts here to choosing a sequence of points  $\mathbf{x}_{n+1}, \dots, \mathbf{x}_{n+N}$  at which the function  $y$  is evaluated. Sequential design decisions, so-called *acquisitions*, are based on the GP model and judiciously balance exploration and exploitation in search for global optima. The GP model is updated after each new value is calculated.

In the noiseless setting ( $\tau = 0$ ), the canonical acquisition function is *expected improvement* (EI) [13]. Define  $f_{\min} = \min_{i=1, \dots, n} y_i$ , the smallest  $y$ -value seen so

far, and let  $I(\mathbf{x}) = \max\{0, f_{\min} - Y(x)\}$  be the *improvement* at  $x$ .  $I(x)$  is largest when  $Y(\mathbf{x})$  has substantial distribution below  $f_{\min}$ . The expectation of  $I(x)$  over  $Y(x)$  has a convenient closed form, revealing balance between exploitation ( $\mu(x)$  under  $f_{\min}$ ) and exploration (large  $\sigma^n(x)$ ):

$$\mathbb{E}\{I(x)\} = (f_{\min} - \mu(x))\Phi\left(\frac{f_{\min} - \mu(x)}{\sigma(x)}\right) + \sigma(x)\phi\left(\frac{f_{\min} - \mu(x)}{\sigma(x)}\right), \quad (3)$$

where  $\Phi$  ( $\phi$ ) is the standard normal cdf (and pdf respectively).

### 3.2 Bayesian optimization of stochastic simulators

When  $y$  is only available through noisy evaluations, the EI acquisition cannot be used directly. Several authors have tackled this issue; we refer to [21] for a review on the topic. We chose here to focus on the *reinterpolation method* proposed in [11], which is based on the use of an instrumental noiseless kriging model, built from the original one. First, the (noisy) kriging predictions at the DOE points  $\mu(\mathbf{x}_1), \dots, \mu(\mathbf{x}_n)$  are computed. Then, a reinterpolating model is built, by using the same covariance kernel and parameters and the same experimental design, but the observation vector is replaced by  $\mu(\mathbf{x}_1), \dots, \mu(\mathbf{x}_n)$  and the noise variance is set to zero. Since this latter model is noise-free, the classical EI can be used as the infill criterion. Once the new design is chosen and the evaluation is performed, both kriging models are updated.

### 3.3 Bayesian optimization with invariances

#### 3.3.1 Definition of local invariances

We first introduce the following notation (this is purely notation, no actual permutation is performed):

$$y(\mathbf{x}) = y(x_i, \mathbf{x}_J, \mathbf{x}_{-iJ}) \quad (4)$$

$$\mathbb{X} = \{\mathbb{X}_i, \mathbb{X}_J, \mathbb{X}_{-iJ}\} \quad (5)$$

**Definition 1 (Simple)** We call *simple invariance* the following case:  $y$  is invariant with respect to  $\mathbf{x}_J$  ( $J$  a subset of  $\{1, \dots, d\} \setminus i$ ) if  $x_i = c_i$  ( $i \in \{1, \dots, d\}$ ):

$$y(c_i, \mathbf{x}_J, \mathbf{x}_{-iJ}) = y(c_i, \mathbf{x}'_J, \mathbf{x}_{-iJ}), \quad \forall \mathbf{x}_J, \mathbf{x}'_J \in \mathbb{X}_J, \mathbf{x}_{-iJ} \in \mathbb{X}_{-iJ}.$$

This corresponds for instance to the last line of Table 2: the observation frequency  $\eta_f$  does not have an effect on the model if the duration of observation  $\gamma_O$  is set to zero.

**Definition 2 (Or)** We call “*or*” *invariance* the following case:  $y$  is invariant with respect to  $\mathbf{x}_J$  ( $J$  a subset of  $\{1, \dots, d\} \setminus I$ ) if there exists at least one  $i \in I$  such that  $x_i = c_i$  ( $I$  a subset of  $\{1, \dots, d\} \setminus J$ ):

$$y(c_i, \mathbf{x}_{I \setminus i}, \mathbf{x}_J, \mathbf{x}_{-IJ}) = y(c_i, \mathbf{x}_{I \setminus i}, \mathbf{x}'_J, \mathbf{x}_{-IJ}), \quad \forall \mathbf{x}_J, \mathbf{x}'_J \in \mathbb{X}_J, \mathbf{x}_{I \setminus i} \in \mathbb{X}_{I \setminus i}.$$

This corresponds for instance to the first line of Table 2: the contamination threshold in the observation zone  $\chi_o$  does not have an effect on the model if the duration of observation  $\gamma_O$  is set to zero or if the probability of detection  $\rho$  is set to zero.

**Definition 3 (Linear)** We call *linear invariance* the following case:  $y$  is invariant with respect to  $\mathbf{x}_J$  ( $J$  a subset of  $\{1, \dots, d\} \setminus I$ ) if  $\mathbf{A}\mathbf{x}_I = \mathbf{b}$ , with  $I$  a subset of  $\{1, \dots, d\} \setminus J$ ,  $\mathbf{A}$  a matrix of size  $p \times \text{Card}(I)$  and  $\mathbf{b}$  a vector of size  $p$ :

$$y(\mathbf{x}_I, \mathbf{x}_J) = y(\mathbf{x}_I, \mathbf{x}'_J), \quad \forall \mathbf{x}_J, \mathbf{x}'_J \in \mathbb{X}_J, \text{ if } \mathbf{A}\mathbf{x}_I = \mathbf{b}.$$

There are two particular cases worth noting:

- setting  $p = \text{Card}(I)$ ,  $\mathbf{A} = \mathbb{I}_p$  and  $\mathbf{b} = \mathbf{c}_I$  results in an “AND” condition:  $y$  is invariant with respect to  $\mathbf{x}_j$  if,  $\forall i \in I, x_i = c_i$ ;
- setting  $p = 1$ ,  $\mathbf{A} = [1, -1]$  results in an invariance under the condition  $x_{i1} = x_{i2}$ .

This invariance case is not illustrated in this work with the sharka problem optimization presented here (with 10 management parameters). However, we may have this situation if we use all the parameters implemented in the model. For instance, a parameter  $\gamma_y$  (not used here) is implemented in the model. It corresponds to the duration of an observation zone for young orchards. In this case, the radius of observation epicenter  $\zeta_{eO}$  does not have an effect on the model if the duration of observation zones  $\gamma_O$  is set to 0 AND if the duration of an observation zone for young orchards  $\gamma_y$  is also set to 0.

### 3.3.2 Principle of input warping

The standard use of GPs implies a hypothesis of stationarity (the unconditional joint probability distribution of the process does not change when shifted in the  $\mathbb{X}$  space), which is in contradiction with the notion of local invariance. *Cette info ne devrait-elle pas etre amenee en amont pour justifier l'intret de la question traitee ?*

There are several ways of incorporating structural information into Gaussian processes. One is to work on the kernel function  $k$  [8, 6]. Another, which is the one we use here, is to transform the original input space  $\mathbb{X}$  into a *warped* one  $\tilde{\mathbb{X}}$  and index the GP on  $\tilde{\mathbb{X}}$ , so that the new topology directly reflects the structural information [32, 15].

Consider for simplicity a single invariance over  $x_J$  when  $x_i = c_i$ . A simple way to handle this problem is to distort locally the space so that the subspace  $\{(x_i, \mathbf{x}_J) | x_i = c_i\}$  collapses to a single point, for instance with  $\mathbf{x}_J$  at its average value:  $(c_i, \bar{\mathbf{x}}_J)$ .

Hence, we are seeking warping functions of the form:

$$\begin{aligned} \psi : \mathbb{X} &\rightarrow \tilde{\mathbb{X}} \\ \mathbf{x} &\mapsto \tilde{\mathbf{x}} = \psi(\mathbf{x}) \end{aligned}$$

such that:

1.  $\psi(x_i, \mathbf{x}_J, \mathbf{x}_{-iJ}) = (c_i, \bar{\mathbf{x}}_J, \mathbf{x}_{-iJ})$  if and only if  $x_i = c_i$ ;
2.  $\psi$  restricted to  $\mathbb{X} \setminus (c_i, ., .)$  and  $\tilde{\mathbb{X}} \setminus (c_i, \bar{\mathbf{x}}_J, .)$  is a diffeomorphism.

In addition, we will search for deformations that decrease monotonically when  $|x_i - c_i|$  increases, that is:

$$\begin{aligned} ((x_i, \mathbf{x}_J, \mathbf{x}_{-iJ}), \psi[(x_i, \mathbf{x}_J, \mathbf{x}_{-iJ})]) &\leq d((x'_i, \mathbf{x}_J, \mathbf{x}_{-iJ}), \psi[(x'_i, \mathbf{x}_J, \mathbf{x}_{-iJ})]) \\ &\quad \text{if } |x_i - c_i| \leq |x'_i - c_i|, \end{aligned}$$

for some distance  $d(., .)$ .

Since the  $\mathbf{x}_J$  dimension collapses to  $\bar{\mathbf{x}}_J$  at  $x_i = c_i$ , we write:

$$\forall j \in J, \quad \tilde{x}_j = \bar{x}_j + (x_j - \bar{x}_j) \alpha(x_i, c_i), \quad (6)$$

with  $\alpha(x_i, c_i)$  an attenuation function such that:

1.  $\alpha(c_i, c_i) = 0$ ;
2.  $\alpha$  increases monotonically with  $|x_i - c_i|$ ;
3.  $0 < \alpha \leq 1, \forall x_i \neq c_i$ .

Condition 1 ensures that  $\tilde{\mathbf{x}}_j = \bar{\mathbf{x}}_j$  when  $x_i = c_i$  (all the dimensions in  $J$  collapse).

### 3.3.3 Warping for a simple invariance

In the simple invariance case, we propose linear and correlation-based attenuation functions:

$$\alpha_{\text{lin}}(x_i, c_i) = \frac{|x_i - c_i|}{\delta_i}, \quad (7)$$

$$\alpha_{\text{cor}}(x_i, c_i) = 1 - r(x_i, c_i), \quad (8)$$

where  $r$  is a  $\mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$  correlation function. Typically,  $\delta_i$  may be set to the range of variation of  $x_i$ , so that the condition  $\alpha \leq 1$  is ensured. Choosing  $r$  as the generalized exponential correlation, we have:

$$\alpha_{\text{exp}}(x_i, c_i) = 1 - \exp \left[ - \left( \frac{|x_i - c_i|}{\theta_i} \right)^d \right], \quad (9)$$

with  $\theta_i$  and  $d$  positive parameters to be tuned.

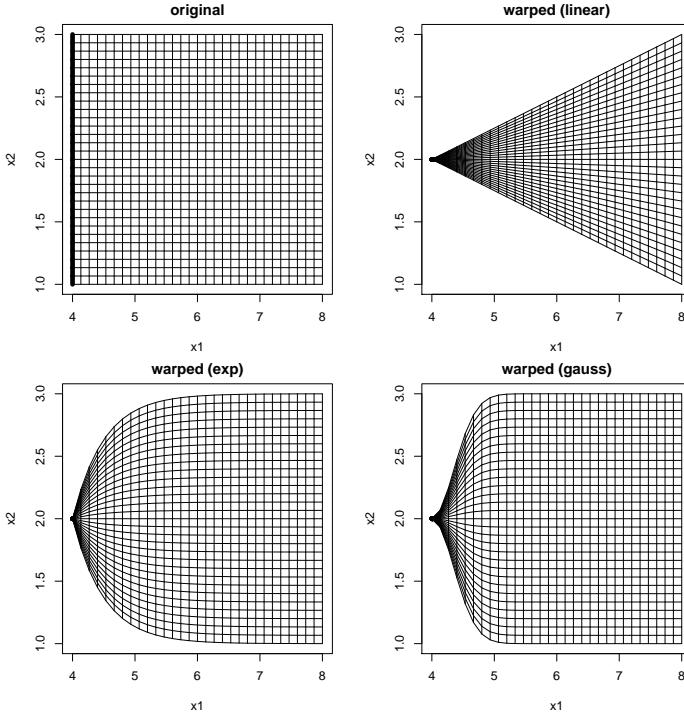
Figure 2 shows a 2D rectangular space distorted by three warpings, when the invariance is on a boundary of  $x_1$ . Figure 3 shows (unconditional) realizations of GPs with a Gaussian kernel applied on the warped space. We see that the invariance at  $x_1$  maximum is ensured. The linear warping induces a strong anisotropy, while with the two other warpings, the process seems stationary far from the critical value.

### 3.3.4 Warping for linear invariances

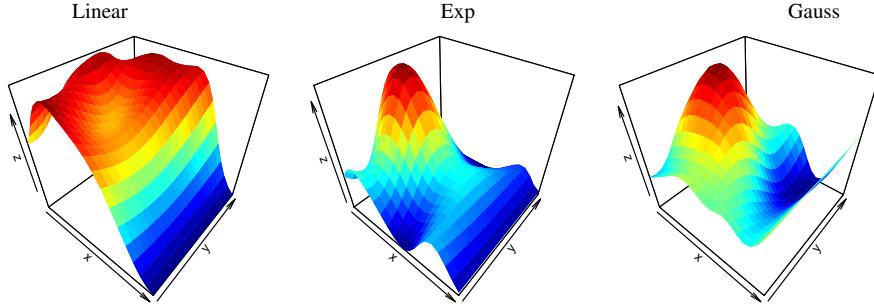
For simplicity, we consider first the particular linear case where  $\mathbf{A} = \mathbb{I}_p$  and  $\mathbf{b} = \mathbf{c}_I$ , that is where invariances occur when a set of variables takes simultaneously a set of critical values:  $\mathbf{Ax}_I = \mathbf{b}$ , or equivalently  $\mathbf{x}_I = \mathbf{c}_I$ . In that case, a possible warping is:

$$\forall j \in J, \quad \tilde{x}_j = \bar{x}_j + (x_j - \bar{x}_j) \alpha_I(\mathbf{x}_I, \mathbf{c}_I). \quad (10)$$

with  $\alpha_I$  now a multivariate attenuation function ( $\mathbb{R}^{\text{Card}(I)} \times \mathbb{R}^{\text{Card}(I)} \rightarrow \mathbb{R}$ ), so that, similarly to the simple case:



**Fig. 2** Three deformations of a 2D space. The local invariance is at  $x_1 = 0$ , highlighted with larger lines.



**Fig. 3** Three GP realizations using warping functions as shown previously.

1.  $\alpha_I(\mathbf{c}_I, \mathbf{c}_I) = 0$ ;
2.  $\alpha_I$  increases monotonically with  $d(\mathbf{x}_I, \mathbf{c}_I)$  (for some distance  $d(., .)$ );
3.  $0 < \alpha_I \leq 1, \forall \mathbf{x}_I \neq \mathbf{c}_I$ .

As in the simple case, linear and correlation-based warpings can be defined as:

$$\alpha_{\text{lin}}(\mathbf{x}_I, \mathbf{c}_I) = \frac{1}{\text{Card}(I)} \sum_{i \in I} \frac{|x_i - c_i|}{\delta_i}, \quad (11)$$

$$\alpha_{\text{cor}}(\mathbf{x}_I, \mathbf{c}_I) = 1 - r_I(\mathbf{x}_I, \mathbf{c}_I), \quad (12)$$

with  $r_I$  a  $\mathbb{R}^{\text{Card}(I)} \times \mathbb{R}^{\text{Card}(I)} \rightarrow \mathbb{R}$  correlation function as in 9.

Generalizing to the affine case  $\mathbf{Ax}_I = \mathbf{b}$ , the warping function is the same as in Equation 10, with now:

$$\alpha(\mathbf{x}_I, \mathbf{c}_I) = 1 - r_A(\mathbf{Ax}_I, \mathbf{b}). \quad (13)$$

### 3.3.5 Combining warpings

*Independent conditions* Now, we consider that we have a series of invariance conditions, defined with respect to sets  $I_1, \dots, I_n$  and corresponding  $J_1, \dots, J_n$ . If  $J_k \cap J_l = \emptyset$ ,  $1 \leq j \neq k \leq n$  and  $I_i \cap J_k = \emptyset$ ,  $1 \leq j, k \leq n$ , the set of warped variables are distinct from the set on which the conditions are written, the invariance conditions are written only once for each variable. In that case, the warpings can be applied independently.

*Combinations of simple conditions: “OR” invariance* Now, we consider the case when  $y$  is invariant w.r.t. a set  $\mathbf{x}_J$  for different conditions on sets  $I_1, \dots, I_n$  (that, for  $\mathbf{x}_{I_1} = \mathbf{c}_{I_1}$  OR  $\mathbf{x}_{I_2} = \mathbf{c}_{I_2}$  OR ...). If  $J \cap I_i = \emptyset$ ,  $1 \leq i \leq n$ , the warping function we propose is:

$$\forall j \in J, \quad \tilde{x}_j = \bar{x}_j + (x_j - \bar{x}_j) \prod_{I \in \{I_1, \dots, I_n\}} \alpha_I(\mathbf{x}_I, \mathbf{c}_I). \quad (14)$$

We see directly that the product of  $\alpha$ 's ensure that  $\tilde{x}_j = \bar{x}_j$  if any  $x_i = c_i$ , and the distortion reduces only when *all* the  $x_i$ 's are far from the  $c_i$ 's.

*“Circular” conditions* Difficulty only arises when some variables appear in both  $I_l$ 's and  $J_m$ 's sets. Take for instance a “reciprocal” condition, e.g.,  $y$  is invariant w.r.t.  $\mathbf{x}_J$  when  $\mathbf{x}_I = \mathbf{c}_I$ , and invariant w.r.t.  $\mathbf{x}_I$  when  $\mathbf{x}_J = \mathbf{c}_J$ . In that case, applying independently warping functions would lead to:

$$\begin{aligned} \psi(\mathbf{c}_I, \mathbf{x}_J, \mathbf{x}_{-IJ}) &= (\mathbf{c}_I, \bar{\mathbf{x}}_J, \mathbf{x}_{-IJ}), \\ \psi(\mathbf{x}_I, \mathbf{c}_J, \mathbf{x}_{-IJ}) &= (\bar{\mathbf{x}}_I, \mathbf{c}_J, \mathbf{x}_{-IJ}), \\ \text{but: } \psi(\mathbf{c}_I, \mathbf{c}_J, \mathbf{x}_{-IJ}) &= (\mathbf{c}_I, \mathbf{c}_J, \mathbf{x}_{-IJ}), \end{aligned}$$

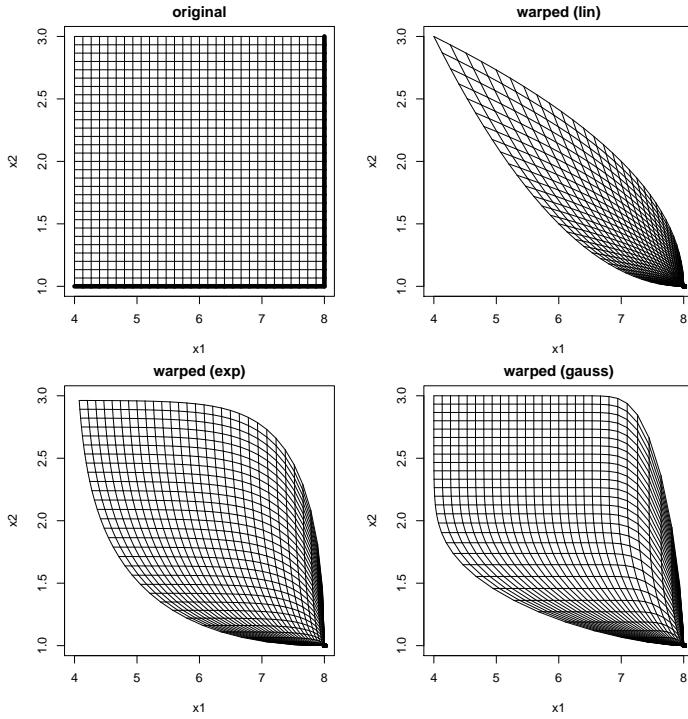
which induces a discontinuity.

In that case, a simple solution is to fix the non influent variable to its critical value instead of its average, hence applying:

$$\forall k \in K = (\cup_{1 \leq l \leq n} I_l) \cap (\cup_{1 \leq m \leq n} J_m), \quad \tilde{x}_k = c_k + (x_k - c_k) \prod_{i \in I_k} \alpha(x_i, c_i) \quad (15)$$

*Remark* This formula does not apply in the affine case (Equation 13).

We show the deformations on a 2D space on Figure 4, where the two critical values are on the boundaries of  $x_1$  and  $x_2$ . Here, the warping of Equation 15 is applied on each variable ( $K = \{1, 2\}$ ). Again, except for the linear warping, the local topology is preserved far from the critical edges.



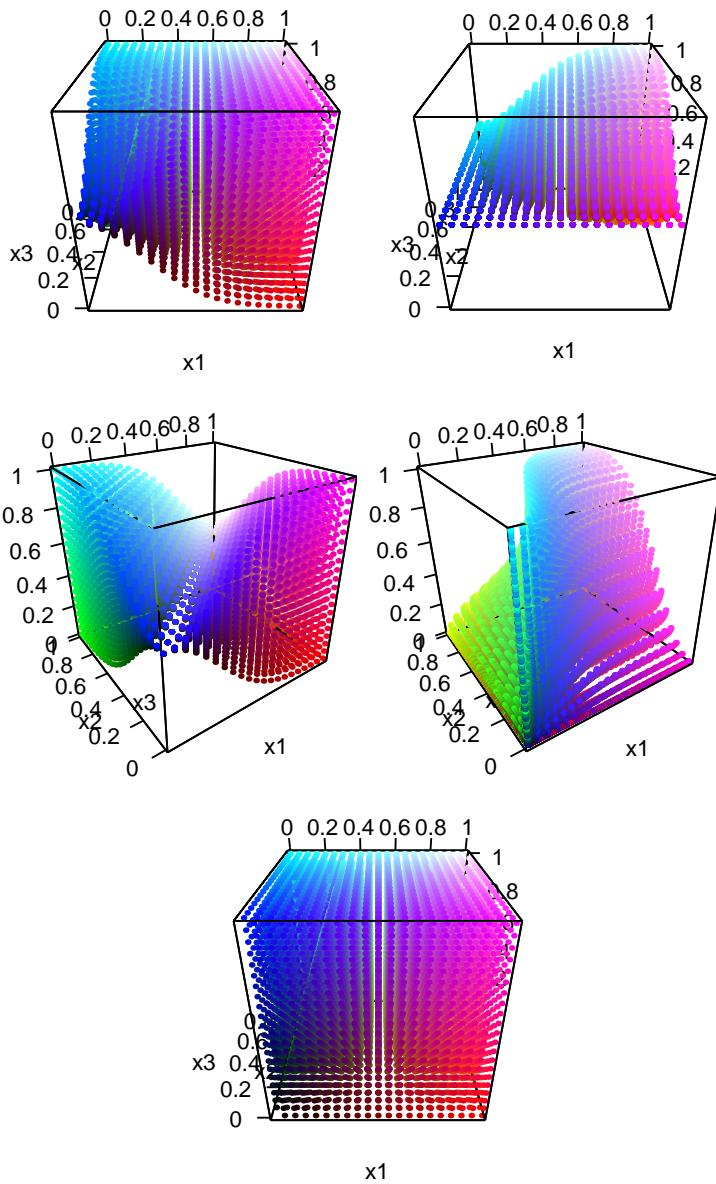
**Fig. 4** Three deformations of a 2D space, with invariance at  $x_1 = 8$  OR  $x_2 = 1$ , highlighted with larger lines.

### 3.3.6 Illustration

Finally, Figure 5 shows four deformations of the unit cubic space, for each of the following invariances:

- AND:  $y$  is invariant w.r.t.  $x_3$  when  $x_1$  AND  $x_2$  are equal to zero (equation 10 with  $I = 1$  and  $J = \{1, 2\}$ );
- OR:  $y$  is invariant w.r.t.  $x_3$  when  $x_1$  OR  $x_2$  are equal to zero (equation 14 with  $I = 1$  and  $J = \{1, 2\}$ );
- LINEAR:  $y$  is invariant w.r.t.  $x_3$  when  $x_1 = x_2 = 0$  (equation 13 with  $I = 1$ ,  $J = \{1, 2\}$ ,  $\mathbf{A} = [1, -1]$  and  $\mathbf{b} = 0$ );
- CIRCULAR:  $y$  is invariant w.r.t. a-  $x_2$  if  $x_1 = 0$ , b-  $x_3$  if  $x_2 = 0$ , c- w.r.t.  $x_1$  if  $x_3 = 0$  (equation 15 with  $K = \{1, 2, 3\}$ ,  $C = [0, 0, 0]$  and  $I_1 = 3$ ,  $I_2 = 1$ , and  $I_3 = 2$ ).

On all cases, a Gaussian warping (exponential with  $d = 2$ ) is applied, with range parameter  $\theta = 0.3$ .



**Fig. 5** Four deformations of the unit cube under different invariances: AND (top left), OR (top right), LINEAR (middle left), CIRCULAR (middle right). The bottom figure shows the orginal space.

### 3.3.7 Warping parameters tuning

The linear warping has the advantage of being parameter-free, which comes at a price of a profound modification of the problem topology. The correlation-based warpings

have the capability of creating more localized distortions, but depend on range parameters (the  $\theta_i$ 's in Equation 9). Those may be estimated by likelihood maximization along with the GP covariance parameters [32, 15].

However, we found in our numerical experiments that choosing the same correlation function for the GP and the warping, and fixing the warping ranges to be 1/10th of the GP ones provided very satisfactory results, while avoiding the extra computational burden.

Note that in the case of linear invariances, choosing the range of the correlation  $r_A$  is non-trivial, as it is not directly linked to design variables. A possible solution is  $\theta_A = \mathbf{A}^T \boldsymbol{\theta}_I$ .

## 4 A warping-based Bayesian optimization of the Sharka model

### 4.1 Numerical setup

#### 4.1.1 Experiments description

To assess the benefits of including the warping step in the optimization process (i.e. reducing the parameter space removing the combinations which lead to the same management), we conducted 50 independent optimizations of sharka management parameters with and without the warping step. Warping is applied to seven variables, following Table 2, to account for two simple invariances  $\eta_s, \eta_f$ , two combined ones  $\chi_o, \eta_{f*}$ , and three implying “circular” conditions:  $\zeta_{eo}, \zeta_f$  and  $\zeta_s$ . On all cases, we used a Matérn 5/2 correlation-based warping.

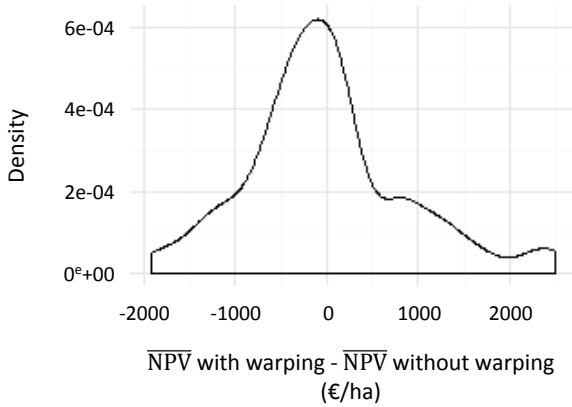
The economic criterion to optimize was the mean of the NPV ( $\overline{NPV}$ ). For this to happen, we randomly selected 50 times 200 management strategies using a maximin Latin hypercube sampling design [7]. Then, for each sampling design of 200 strategies, we performed 2 optimizations in parallel: with and without the warping step. For each optimization, we performed sequentially 200 iterations allowing to choose 200 new strategies, resulting in a total of 400 evaluated strategies. For each evaluated strategy, the objective function is computed by averaging over 1,000 simulations (carried out with different random seeds) to take into account the variability due to the epidemic and landscape characteristics.

#### 4.1.2 Bayesian optimization setup

For all experiments, we used the same GP modeling setup, that is, an unknown constant trend (ordinary kriging, [16]) and Matérn 5/2 covariance function [24, Chapter 4]. The acquisition function maximized at each step is the expected improvement on the *reinterpolating* model. The maximization is performed by a large-scale random search followed by a local optimization starting for the optimum found by the random search (i.e. the evaluated points in the optimization process are chosen around the best current  $\overline{NPV}$ ) All experiments were conducted in R [23], using code adapted from the `DiceOptim` package [20].

## 4.2 Results

We firstly compared the optimization results by subtracting the  $\overline{NPV}$  achieved using the optimization with the warping step and the optimization without the warping step (obtained from the same sampling design). In 24 out of the 50 optimization cases, we obtained better  $\overline{NPV}$  with the warping step than without. This point is illustrated by the probability density function which is centered on 0 (Fig.6). This result means that with 200 iterations in the optimization, the final optimization result is not impacted by the use of the warping.



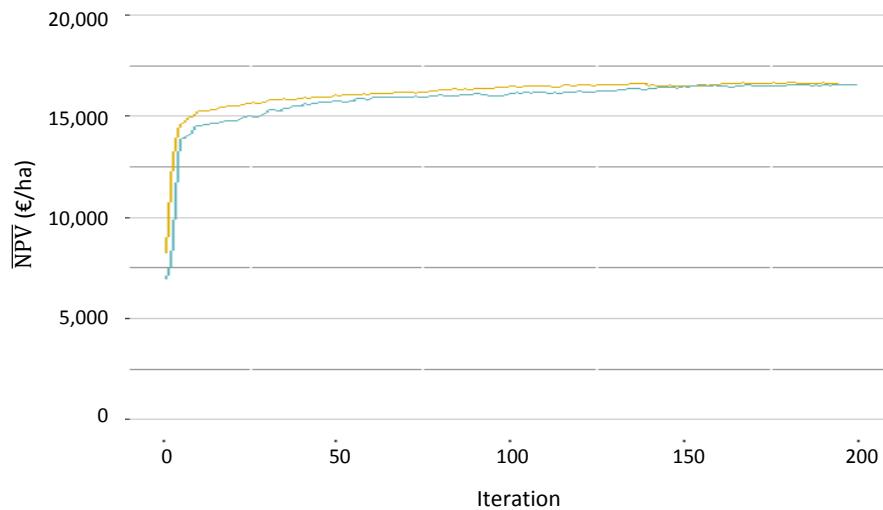
**Fig. 6** Comparison of  $\overline{NPV}$  obtained at the end of the optimization with and without warping.

However, we showed that the warping can impact the optimization speed (Fig.7). Indeed, at the 3<sup>rd</sup> iteration, the gap between the yellow (with warping) and the blue (without warping) lines is already 3957 euro/ha. In addition, to reach  $\overline{NPV}=16,400$  euro/ha, we needed in average only 96 iterations in the optimization process with warping against 144 iterations without warping.

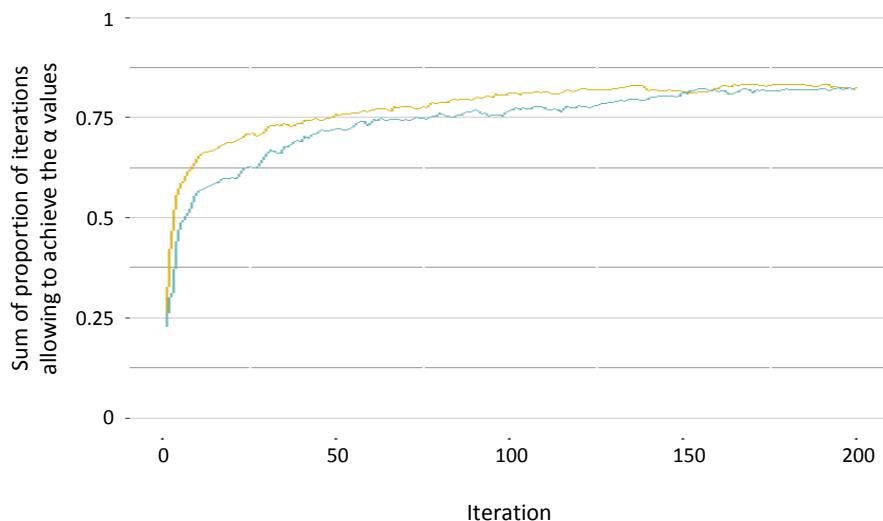
To go further, we performed a nonlinear regression of  $\overline{NPV}$  obtained for all the selected strategies during the optimization process with and without the warping step, and we compared the growth parameter  $c$  of the following regression:  $\overline{NPV} = A + be^{-cx_i}$ . This parameter was higher with (0.26) than without (0.18) warping.

In addition, we can visually observe that the warping step allow to improve the optimization speed on the Fig.8 and Supplementary Fig.1. These figures were represented with a specific algorithm based on empirical distribution functions [10]. Briefly, we uniformly defined 100  $\alpha$  values within a specified range. Then, for each iteration performed in the optimization process (i.e. for each of the 200 evaluated strategies), we add: the number of optimizations (among 50) which exceed  $\alpha_1$ , the number of optimizations which exceed  $\alpha_2$ , ..., the number of optimizations which exceed  $\alpha_{100}$ . We used  $\alpha \in [0;18,012.12]$  Supplementary (Fig.1) and  $\alpha \in [10,000;18,012.12]$

(Fig.8). The value 18,012.12 corresponds to the maximal value of  $\overline{NPV}$  identified in all the optimizations.



**Fig. 7** Comparison of  $\overline{NPV}$  obtained during optimizations with and without warping. Yellow and blue lines represent the mean of the  $\overline{NPV}$  selected at each iteration for the 50 optimizations respectively performed with and without the warping step.



**Fig. 8** Results of the algorithm using empirical distribution functions [10] with (yellow) and without (blue) warping ( $\alpha \in [10,000; 18,012.12]$ ).

## 5 Conclusion

In this study, we showed how a Bayesian optimization process can be improved by accounting for some prior structural information: the insensitivity of the model with respect to a subset of its input variables when another subset of inputs takes a particular value. Such *local invariances* were exhibited by our spatiotemporal model simulating sharka management, characterized by 10 parameters related to the surveillance of the orchards. In this example, the invariances arise because parameters (radius of different zones, surveillance frequency in each zone, detection probability of infected trees, and duration of observation zones) are strongly related. Indeed, we easily note, for instance, that when the detection probability takes a value of 0, numerous other parameters do not influence the model results.

To tackle this problem, we proposed to use a warping of the input space, that here amounted to remove locally dimensions of the input space. The warping we used is based on correlation functions, making it very simple to implement while allowing sufficient flexibility. A particular advantage of input warping over other approaches is that it can be straightforwardly embedded in a BO algorithm. *ne faudrait-il pas développer un peu cet argument qui permet d'indiquer la plus-value de l'approche par rapport ce qui existe déjà ?*

We applied this Bayesian optimization process to the spatio-temporal sharka model. We performed various optimizations of its management parameters firstly with the use of warping (which allows accounting for the invariances) and then without. We showed that both approaches led to the same maximal  $\overline{NPV}$ , but the the optimization process with warping was substantially faster, showing that the warping efficiently reduced the search space without altering the exploration / exploitation trade-off.

As future steps for this research, we could first embed learning the warping parameters together with the parameters of the GP covariance in a single likelihood maximization step. Another room for improvement is to adapt the EI maximization step to the new topology induced by the warping (here, on all experiments the EI was maximized over the original space). Finally, the optimization strategy pursued here used a large fixed number of replicates (1,000) for each evaluated design. Combining warping with an efficient adaptative scheme to handle replicates [12] would drastically reduce the cost of the optimization.

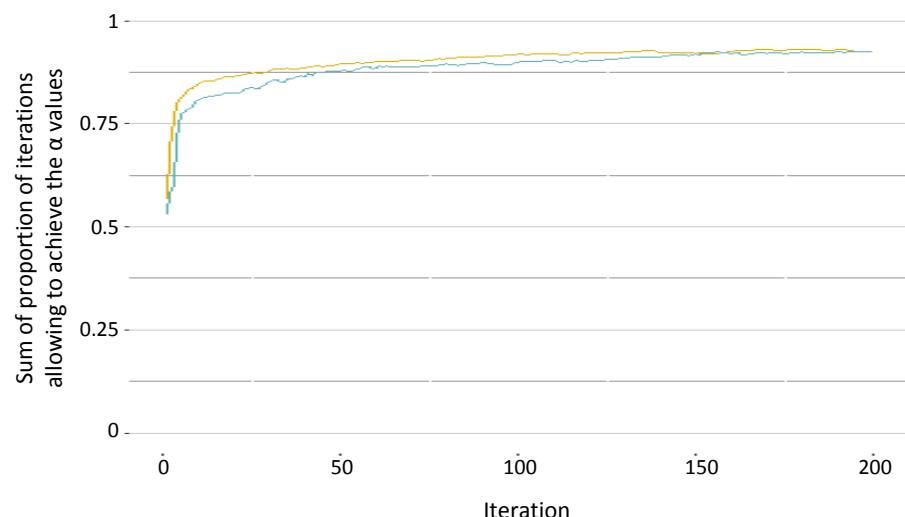
## References

1. Bajardi P, Barrat A, Savini L, Colizza V (2012) Optimizing surveillance for live-stock disease spreading through animal movements. *Journal of the Royal Society Interface* 9(76):2814–2825
2. Box GE, Draper NR (1987) Empirical model-building and response surfaces. John Wiley & Sons
3. Cambra M, Capote N, Myrta A, Llácer G (2006) Plum pox virus and the estimated costs associated with sharka disease. *EPPO Bulletin* 36(2):202–204
4. Cunniffe NJ, Koskella B, Metcalf CJ, Parnell S, Gottwald TR, Gilligan CA (2015) Thirteen challenges in modelling plant diseases. *Epidemics* 10:6–10

5. Cunniffe NJ, Cobb RC, Meentemeyer RK, Rizzo DM, Gilligan CA (2016) Modeling when, where, and how to manage a forest epidemic, motivated by sudden oak death in California. *Proceedings of the National Academy of Sciences of the United States of America* 113(20):5640–5645
6. Duvenaud D (2014) Automatic model construction with gaussian processes. PhD thesis, University of Cambridge
7. Fang KT, Li R, Sudjianto A (2005) Design and modeling for computer experiments. Chapman and Hall/CRC
8. Ginsbourger D, Durrande N, Roustant O (2013) Kernels and designs for modelling invariant functions: From group invariance to additivity. In: mODa 10—Advances in Model-Oriented Design and Analysis, Springer, pp 107–115
9. Grechi I, Ould-Sidi MM, Hilgert N, Senoussi R, Sauphanor B, Lescourret F (2012) Designing integrated management scenarios using simulation-based and multi-objective optimization: Application to the peach tree—*myzus persicae* aphid system. *Ecological Modelling* 246:47–59
10. Hansen N, Auger A, Ros R, Finck S, P P (2010) Comparing results of 31 algorithms from the black-box optimization benchmarking bbob-2009. *Proceedings of the 12th annual conference companion on Genetic and evolutionary computation* pp 1689–1696
11. J Forrester AI, Keane AJ, Bressloff NW (2006) Design and analysis of “noisy” computer experiments. *AIAA journal* 44(10):2331–2339
12. Jalali H, Van Nieuwenhuyse I, Picheny V (2017) Comparison of kriging-based algorithms for simulation optimization with heterogeneous noise. *European Journal of Operational Research* 261(1):279–301
13. Jones DR, Schonlau M, Welch WJ (1998) Efficient global optimization of expensive black-box functions. *Journal of Global Optimization* 13(4):455–492
14. Kompas T, Ha PV, Nguyen HTM, East I, Roche S, Garner G (2017) Optimal surveillance against foot-and-mouth disease: the case of bulk milk testing in Australia. *Australian Journal of Agricultural and Resource Economics* 61(4):515–538
15. Marmin S, Ginsbourger D, Baccou J, Liandrat J (2018) Warped gaussian processes and derivative-based sequential designs for functions with heterogeneous variations. *SIAM/ASA Journal on Uncertainty Quantification* 6(3):991–1018
16. Matheron G (1963) Principles of geostatistics. *Economic Geology* 58(8):1246–1266
17. Mockus J (2012) Bayesian approach to global optimization: theory and applications, vol 37. Springer Science & Business Media
18. Mushayabasa S, Tapedzesa G (2015) Modeling the effects of multiple intervention strategies on controlling foot-and-mouth disease. *BioMed research international*
19. Picard C, Soubeyrand S, Jacquot E, Thébaud G (2018) Analyzing the influence of landscape aggregation on disease spread to improve management strategies. (under review)
20. Picheny V, Ginsbourger D (2014) Noisy kriging-based optimization methods: a unified implementation within the DiceOptim package. *Computational Statistics & Data Analysis* 71:1035–1053

21. Picheny V, Wagner T, Ginsbourger D (2013) A benchmark of kriging-based infill criteria for noisy optimization. *Structural and Multidisciplinary Optimization* 48(3):607–626
22. Pleydell DRJ, Soubeyrand S, Dallot S, Labonne G, Chadoeuf J, Jacquot E, Thébaud G (2018) Estimation of the dispersal distances of an aphid-borne virus in a patchy landscape. *PLoS Computational Biology* 14(4):e1006085
23. R Core Team (2018) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL <https://www.R-project.org/>
24. Rasmussen CE, Williams C (2006) Gaussian Processes for Machine Learning. MIT Press, URL <http://www.gaussianprocess.org/gpml/>
25. Rimbaud L, Dallot S, Gottwald T, Decroocq V, Jacquot E, Soubeyrand S, Thébaud G (2015) Sharka epidemiology and worldwide management strategies: learning lessons to optimize disease control in perennial plants. *Annual Review of Phytopathology* 53:357–378
26. Rimbaud L, Bruchou C, Dallot S, Pleydell DRJ, Jacquot E, Soubeyrand S, Thébaud G (2018) Using sensitivity analysis to identify key factors for the propagation of a plant epidemic. *Royal Society Open Science* 5(1):171435
27. Rimbaud L, Dallot S, Bruchou C, Thoyer S, Jacquot E, Soubeyrand S, Thébaud G (2018) Heuristic optimisation of the management strategy of a plant epidemic using sequential sensitivity analyses. *BioRxiv* 315747
28. Rios LM, Sahinidis NV (2013) Derivative-free optimization: a review of algorithms and comparison of software implementations. *Journal of Global Optimization* 56(3):1247–1293
29. Roustant O, Ginsbourger D, Deville Y (2012) DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based meta-modeling and optimization. *Journal of Statistical Software* 51(1):1–55, URL <http://www.jstatsoft.org/v51/i01/>
30. Shahriari B, Swersky K, Wang Z, Adams RP, De Freitas N (2016) Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE* 104(1):148–175
31. Snelson E, Ghahramani Z, Rasmussen CE (2004) Warped gaussian processes. In: *Advances in neural information processing systems*, pp 337–344
32. Snoek J, Swersky K, Zemel R, Adams R (2014) Input warping for bayesian optimization of non-stationary functions. In: *International Conference on Machine Learning*, pp 1674–1682
33. Tildesley MJ, Savill NJ, Shaw DJ, Deardon R, Brooks SP, Woolhouse ME, Grenfell BT, Keeling MJ (2006) Optimal reactive vaccination strategies for a foot-and-mouth outbreak in the UK. *Nature* 440(7080):83
34. VanderWaal K, Enns EA, Picasso C, Alvarez J, Perez A, Fernandez F, Gil A, Craft M, Wells S (2017) Optimal surveillance strategies for bovine tuberculosis in a low-prevalence country. *Scientific Reports* 7(1):4140

## Supporting information



**Supplementary Figure 1** Results of the algorithm using empirical distribution functions [10] with (yellow) and without (blue) warping ( $\alpha \in [0;18,012.12]$ ).