Optimal control strategies for the sterile mosquitoes technique

Luís Almeida, Michel Duprez, Yannick Privat, Nicolas Vauchelet

Optimal control strategies for the sterile mosquitoes technique

Luis Almeida * Michel Duprez† Yannick Privat‡ Nicolas Vauchelet§

November 15, 2021

Abstract

Mosquitoes are responsible for the transmission of many diseases such as dengue fever, zika or chigungunya. One way to control the spread of these diseases is to use the sterile insect technique (SIT), which consists in a massive release of sterilized male mosquitoes. This strategy aims at reducing the total population over time, and has the advantage being specific to the targeted species, unlike the use of pesticides. In this article, we study the optimal release strategies in order to maximize the efficiency of this technique. We consider simplified models that describe the dynamics of eggs, males, females and sterile males in order to optimize the release protocol. We determine in a precise way optimal strategies, which allows us to tackle numerically the underlying optimization problem in a very simple way. We also present some numerical simulations to illustrate our results.

Keywords: Sterile insect technique, population dynamics, optimal control problem, Pontryagin Maximum Principle (PMP).
2010 AMS subject classifications: 92D25, 49K15, 65K10

1 Introduction

The sterile insect technique (SIT) consists in massively releasing sterilized males in the area where one wishes to reduce the development of certain insects (mosquitoes in this case). Since the released sterile males mate with females, the number of offspring is then reduced, and the size of the insect population diminishes. This strategy was first studied by R. Bushland and E. Knipling and applied successfully in the early 1950's by nearly eradicating screw-worm fly in North America. Since then, this technique has been considered for different pests and disease vectors [7, 17].

Among such vectors, mosquitoes (including Aedes mosquitoes) are responsible for the transmission to humans of many diseases for which there is currently no efficient vaccine nor treatment. Thus, the sterile insect technique and the closely related incompatible insect technique are very promising tools to control the spread of such diseases by reducing the size of the vector population (there are cases where these techniques have been successfully used to drastically reduce mosquito populations in some isolated regions, e.g. [28, 31]).

*Sorbonne Université, CNRS, Université de Paris, Inria, Laboratoire Jacques-Louis Lions UMR7598, F-75005 Paris, France (luis.almeida@sorbonne-universite.fr)

†Inria, équipe MIMESIS, Université de Strasbourg, Icube, CNRS UMR 7357, Strasbourg, France (michel.duprez@inria.fr).

‡IRMA, Université de Strasbourg, CNRS UMR 7501, Inria, 7 rue René Descartes, 67084 Strasbourg, France (yannick.privat@unistra.fr).

§Laboratoire Analyse, Géométrie et Applications CNRS UMR 7539, Université Sorbonne Paris Nord, Villetaneuse, France (vauchelet@math.univ-paris13.fr).

In order to study the efficiency of this technique and to optimize it, mathematical modeling is of great use. For instance, in [5, 14, 15, 3], the authors propose mathematical models to study the dynamics of the mosquito population when releasing sterile males. Recently in [29], the authors propose and analyze a differential system modeling the mosquito population dynamics. Their model is based on experimental observations and is constructed by assuming that there is a strong Allee effect in the insect population dynamics. A similar model, without strong Allee effect, is investigated in [4]. Control theory also allows to study the feasibility of controlling the population thanks to the sterile insect technique, and it has been studied in several works, see e.g. [9, 10, 6]. Using such mathematical models, authors are able to compare the impact of different strategies in releasing sterile mosquitoes (see e.g. [12, 25] and [15, 23] where periodic impulsive releases are considered).

In order to find the best possible release protocol, optimal control theory may be used. In [18], optimal control methods are applied to the rate of introduction of sterile mosquitoes. An approach developed in [30] attempts to control both breeding rates and the rate of introduction of sterile mosquitoes. In [19], the influence of habitat modification is also considered. Finally, existence and numerical simulations of the solution to an optimal control problem for the SIT has been proposed in [11].

In this paper, we study how to optimize the release protocol in order to minimize certain cost functionals, such as the number of mosquitoes. Starting with the mathematical model presented in [29] without Allee effect, we investigate some optimal control problems and focus on obtaining a precise description of the optimal control. To do so, we consider a simplified version of the mathematical model and we perform a complete study of the optimizers. In particular, in our main result, we describe precisely the optimal release function to minimize the number of sterile males needed to reach a given size of the population of mosquitoes. Our theoretical results are illustrated with some numerical simulations. We also provide some extensions to certain related optimization problems in order to illustrate the robustness of our approach.

The outline of this paper is as follow. In Section 2, we introduce the mathematical model we will adopt for the sterile insect technique, and describe some useful qualitative properties related to stability issues. For the sake of readability, all the proofs will be postponed to Appendix A. Section 3 is devoted to the introduction of the problems modeling the search of optimal release protocols and the statement of the main theoretical results providing a precise description of optimal strategies. We then derive a simple algorithm to compute them numerically and provide illustrating simulations. The proofs of the main theoretical results are postponed to Section 4. Finally, some comments on other possible approaches are gathered in Section 5.

2 Mathematical modelling

2.1 Mosquito life cycle

The life cycle of a mosquito (male or female) consists of several stages and takes place successively in two distinct environments: it includes an aquatic phase (egg, larva, pupa) and an aerial phase (adult). A few days after mating, a female mosquito may lay a few dozen eggs, possibly spread over several breeding sites. Once laid, the eggs of some species can withstand hostile environments (including adverse weather conditions) for up to several months before hatching. This characteristic contributes to the adaptability of mosquitoes and has enabled them to colonize temperate regions. After stimulation (e.g. rainfall), the eggs hatch to give birth to larvae that develop in the water and reach the pupal state. This larval phase can last from a few days to a few weeks. Then, the insect undergoes its metamorphosis. The pupa (also called *nymph*) remains in the aquatic state for 1 to 3 days and then becomes an adult mosquito (or *imago*): it is the emergence and the beginning of the

aerial phase. The lifespan of an adult mosquito is estimated to be of a few weeks.

In many species, egg laying is only possible after a blood meal, i.e. the female must bite a vertebrate before each egg laying. This behavior, called hematophagy, can be exploited by infectious agents (bacteria, viruses or parasites) to spread, alternately from a vertebrate host (humans, for what we are interested in here) to an arthropod host (here, the mosquito).

Based on these observations, a compartmental model has been introduced in [29] to model the life cycle of mosquitoes when releasing sterile mosquitoes. In what follows, we will both deal with the full and a simplified version of [29]. The reason for studying such a simplified model is twofold: on the one hand, the simplified model can be considered relevant from a biological point of view within certain limits. On the other hand, the study of such a "prototype" model can be considered as a first step towards the development of robust control methodologies with a wider application.

To this aim, we will denote by $u(\cdot)$ a control function standing for a sterile male release function (in other words the rate of sterile male mosquitoes release at each time) and by

- $M_s(t)$, the sterilized adult males at time t;
- $F(t)$, the adult females that have been fertilized at time t.

The system we will use for describing the behavior of the mosquito population under the action of the control $u(\cdot)$ reads

$$\begin{cases} \dfrac{dF}{dt} = f(F, M_s), \\ \dfrac{dM_s}{dt} = u - \delta_s M_s, \end{cases} \tag{$\mathcal{S}_1$}$$

where $f : \mathbb{R}^2 \to \mathbb{R}$ denotes the nonlinear function

$$f(F, M_s) = \frac{\nu(1-\nu)\beta_E^2 \nu_E^2 F^2}{\left(\frac{\beta_E F}{K} + \nu_E + \delta_E\right)\left((1-\nu)\nu_E \beta_E F + \delta_M \gamma_s M_s (\frac{\beta_E F}{K} + \nu_E + \delta_E)\right)} - \delta_F F, \tag{1}$$

with the following parameter choices:

- $\beta_E > 0$ is the oviposition rate;
- $\delta_E, \delta_M, \delta_F, \delta_s > 0$ are the death rates for eggs, adult males, females, and sterile males respectively;
- $\nu_E > 0$ is the hatching rate for eggs;
- $\nu \in (0, 1)$ the probability that a pupa gives rise to a female, and $(1 - \nu)$ is therefore the probability to give rise to a male;
- $K > 0$ is the environmental capacity for eggs. It can be interpreted as the maximum density of eggs that females can lay in breeding sites;
- $\gamma_s > 0$ accounts for the fact that females may have a preference for fertile males. Then, the probability that a female mates with a fertile male is $\frac{M}{M+\gamma_s M_s}$.

In the following section, we explain and comment on the choice of this system.

2.2 Derivation of the simplified model and presentation of the original one

The choice of ($\mathcal{S}_1$) as model is inspired by [29]. To explain how it has been derived, let us present the more involved model we have considered. Let us introduce:

- $E(t)$, the mosquito density in aquatic phase at time t;
- $M(t)$, the adult male density at time t;
- $M_s(t)$, the sterilized adult male density at time t;
- $F(t)$, the density of adult females that has been fertilized at time t.

Then, the dynamics of the mosquito population is driven by the following dynamical system:

$$\begin{cases} \dfrac{dE}{dt} = \beta_E F\left(1 - \dfrac{E}{K}\right) - \big(\nu_E + \delta_E\big)E, \\ \dfrac{dM}{dt} = (1-\nu)\nu_E E - \delta_M M, \\ \dfrac{dF}{dt} = \nu\nu_E E \dfrac{M}{M + \gamma_s M_s} - \delta_F F, \\ \dfrac{dM_s}{dt} = u - \delta_s M_s. \end{cases} \tag{$\mathcal{S}_2$}$$

Regarding this latter model, the main difference with the one in [29] is the absence of an exponential term in the equation on F to introduce an Allee effect. This effect reflects the fact that, when the population density is very low, it can be difficult to find a partner to mate. This term is important when considering a small population size. Here, since we are focusing on large populations that we want to reduce in size, we will neglect this term.

Assuming that the time dynamics of the mosquitoes in aquatic phase and the adult males compartments are fast leads to assume that the equations on $E(\cdot)$ and $M(\cdot)$ are at equilibrium. We refer for instance to [1] for additional explanations on the justification for these asymptotics. Hence, we get the following equalities

$$E = \frac{\beta_E F}{\frac{\beta_E F}{K} + \nu_E + \delta_E} \quad \text{and} \quad M = \frac{(1-\nu)\nu_E}{\delta_M} E \ .$$

Plugging such expressions into ($\mathcal{S}_2$) allows us to obtain ($\mathcal{S}_1$).

We conclude this paragraph by numerically comparing the full model ($\mathcal{S}_2$) and the simplified one ($\mathcal{S}_1$) that we will aim to control. We consider the numerical values taken from [29, Table 3] and recalled in Table 1 below.

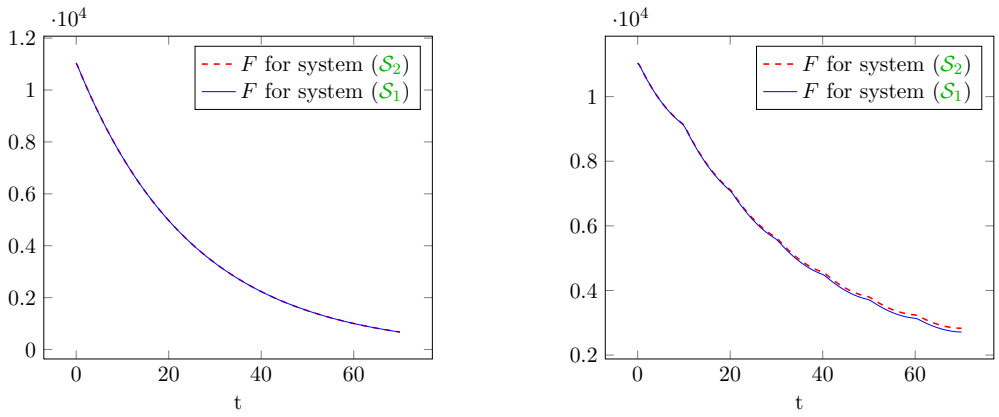

Figure 1: Comparisons of the numerical solutions F (in continous blue line for System ($\mathcal{S}_2$) and in dashed red line for System ($\mathcal{S}_1$) or equation 4). The initial conditions correspond to the "persistence" equilibrium (see propositions 2.1 and 2.3) Left: $u(\cdot) = 15\,000$. Right : the releases occur every 10 days with an intensity of 20 000 mosquitoes in one day, i.e. $u(t) = 20\,000 \sum_{k=0}^{6} \mathbb{1}_{[10k,10k+1]}(t)$.

Parameter	*Name*	*Value interval*	*Chosen value*	*Unit*
β_E	Effective fecundity	7.46–14.85	10	Day^{-1}
γ_s	Mating competitiveness of sterilizing males	0–1	1	-
ν_E	Hatching parameter	0.005–0.25		Day^{-1}
δ_E	Mosquitoes in aquatic phase death rate	0.023 - 0.046	0.03	Day^{-1}
δ_F	Female death rate	0.033 - 0.046	0.04	Day^{-1}
δ_M	Males death rate	0.077 - 0.139	0.1	Day^{-1}
δ_s	Sterilized male death rate		0.12	Day^{-1}
ν	Probability of emergence		0.49	-

Table 1: Value intervals of the parameters for systems ($\mathcal{S}_2$) and ($\mathcal{S}_1$) (see [29])

The numerical results are shown in Fig. 1. In these simulations, the time of the experiment is assumed to be $T = 70$ days, and we choose two different release functions u:

$$u(t) = 15\,000 \text{ (left)}, \quad \text{and} \quad u(t) = 20\,000 \sum_{k=0}^{6} \mathbb{1}_{[10k,10k+1]}(t) \text{ (right)}.$$

In what follows, for a subset A of a given set E, the notation $\mathbb{1}_A$ stands for the characteristic function of A. Namely, for each $x \in E$, $\mathbb{1}_A(x)$ is equal to 1 if $x \in A$ and 0 otherwise. The dynamics for F (female compartment) is represented for both systems ($\mathcal{S}_2$) (blue, continuous line) and ($\mathcal{S}_1$) (red, dashed line). We observe that both problems are very close, which indicates that the dynamics of fertilized females in system ($\mathcal{S}_2$) may be approximated by the one in system ($\mathcal{S}_1$).

A mathematical element of this observation lies in the fact that the equilibria of Systems ($\mathcal{S}_2$) and ($\mathcal{S}_1$) coincide. When dealing with optimal control properties, we will also numerically observe in Section 3.3 that this simplification does not affect the optimal strategies in a strong way.

2.3 Mathematical properties of the dynamical systems

This section is devoted to establishing stability properties for equilibria of Systems ($\mathcal{S}_2$) and ($\mathcal{S}_1$) in the absence of control, in order to qualitatively understand their behavior whenever initial data are chosen close to equilibria. These results can be considered as preliminary tools before investigating optimal control properties for these problems.

Recall that both systems share the same steady states. We will moreover show that they enjoy the same stability properties. For the sake of readability, all proofs are postponed to Appendix A.

In what follows, we will make the following assumption, in accordance with the numerical values gathered in Table 1:

$$\delta_s > \delta_M \quad \text{and} \quad \mathcal{R}_0 := \frac{\nu \beta_E \nu_E}{\delta_F(\nu_E + \delta_E)} > 1, \tag{$\mathcal{H}$}$$

where $\mathcal{R}_0$ denotes the so-called basic offspring number (number of adult females produced by one adult female during her lifespan).

Proposition 2.1 (Stability properties for System ($\mathcal{S}_2$))**.** *Let us assume that ($\mathcal{H}$) holds.*

(i) If $u(\cdot) = 0$, then, System ($\mathcal{S}_2$) has two equilibria:

- *the "extinction" equilibrium $(E_1^*, M_1^*, F_1^*, M_{s1}^*) = (0, 0, 0, 0)$, which is linearly unstable[1];*
- *the "persistence" equilibrium $(E_2^*, M_2^*, F_2^*, M_{s2}^*) = \big(\overline{E}, \overline{M}, \overline{F}, 0\big)$, where*

$$\overline{E} = K\left(1 - \frac{1}{\mathcal{R}_0}\right), \qquad \overline{M} = \frac{(1-\nu)\nu_E}{\delta_M}\overline{E}, \qquad \overline{F} = \frac{\nu\nu_E}{\delta_F}\overline{E}, \tag{2}$$

which is locally asymptotically stable (LAS).

(ii) If the control function u is assumed to be non-negative, then the corresponding solution (E, M, F, M_s) to System ($\mathcal{S}_2$) enjoys the following stability property:

$$\begin{cases} E(0) \in (0, \overline{E}] \\ M(0) \in (0, \overline{M}] \\ F(0) \in (0, \overline{F}] \\ M_s(0) \geqslant 0 \end{cases} \Longrightarrow \begin{cases} E(t) \in (0, \overline{E}] \\ M(t) \in (0, \overline{M}] \\ F(t) \in (0, \overline{F}] \\ M_s(t) \geqslant 0 \end{cases} \quad \text{for all } t \geqslant 0.$$

Finally, let U^* be defined by

$$U^* := \mathcal{R}_0 \frac{K(1-\nu)\nu_E \delta_s}{4\gamma_s \delta_M}\left(1 - \frac{1}{\mathcal{R}_0}\right)^2 \tag{3}$$

and let $\overline{U}$ denote any positive number such that $\overline{U} > U^*$. If $u(\cdot)$ denotes the constant control function almost everywhere equal to $\overline{U}$ for all $t \geqslant 0$, then the corresponding solution $(E(t), M(t), F(t))$ to System ($\mathcal{S}_2$) converges to the extinction equilibrium as $t \to +\infty$.

Remark 2.2. *We verify from the first point of this proposition that $\mathcal{R}_0 > 1$ implies the population persistence while $\mathcal{R}_0 \leqslant 1$ expresses the population extinction.*

In the following result, we will again use the notations introduced in Prop. 2.1 above.

Proposition 2.3 (Stability properties for System ($\mathcal{S}_1$))**.** *Let us assume that ($\mathcal{H}$) holds.*

[1] Meaning that at least one eigenvalue of the Jacobian matrix of the system has a positive real part

(i) If $u(\cdot) = 0$, System $(\mathcal{S}_1)$ has two equilibria:

- *the "extinction" equilibrium $(F_1^*, M_{s1}^*) = (0, 0)$, which is unstable if $\delta_s > \delta_F$.*
- *the "persistence" equilibrium $(F_2^*, M_{s2}^*) = \big(\overline{F}, 0\big)$, is locally asymptotically stable (LAS).*

(ii) If the control function u is assumed to be non-negative, then the corresponding solution (F, M_s) to System $(\mathcal{S}_1)$ enjoys the following stability property:

$$\left\{ \begin{array}{l} F(0) \in (0, \overline{F}] \\ M_s(0) \geqslant 0 \end{array} \right. \Longrightarrow \left\{ \begin{array}{l} F(t) \in (0, \overline{F}] \\ M_s(t) \geqslant 0 \end{array} \right. \quad \text{for all } t \geqslant 0.$$

If $u(\cdot)$ denotes the constant control function almost everywhere equal to $\overline{U} > U^*$ (defined by (3)), then $F(t)$ goes to 0 as $t \to +\infty$.

As a consequence of Propositions 2.1 and 2.3, it follows that if the horizon of time of control T is large enough, by releasing a sufficient amount of mosquitoes, it is possible to make the wild population of mosquitoes be as small as desired at time T. In the next section, we will investigate the issue of minimizing the number of released mosquitoes in order to reach a given size of the wild population. In Section 5, we will also comment on other relevant minimization problems related to the sterile insect technique.

3 Optimal control problems

In what follows, we will consider an initial state corresponding to the beginning of the experiment, in which there are no sterile mosquitoes. Hence, the mosquito population is at the "persistence" equilibrium at the beginning of the experiment, meaning that one chooses

$$\begin{array}{c} F(0) = \overline{F}, \quad M_s(0) = 0, \text{ for System } (\mathcal{S}_2), \text{ to which we should add} \\ E(0) = \overline{E}, \quad M(0) = \overline{M}, \text{ for System } (\mathcal{S}_2) \end{array} \tag{4}$$

where $(\overline{E}, \overline{M}, \overline{F})$ are defined by (2). Our aim is to determine the optimal time distribution of the releases, such that the number of mosquitoes used to reach a desired size of the population is as small as possible.

3.1 Statement of the problems and main results

Let us first model the optimization of the release procedure. Given the duration of the experiment, we aim at minimizing the amount of mosquitoes required to reduce the size of the wild population of mosquitoes to a given value:

What should be the time distribution of optimal releases in order to reach a given value of the wild population at the end of the experiment, by using as few sterilized mosquitoes as possible?

To answer this question, we first define a cost functional that will stand for the objective we are trying to achieve. We mention that other formulations of minimization problem are also proposed in Section 5.

Let $T > 0$ be a given horizon of time, $\overline{U} > 0$ be a maximal amount of sterilized mosquitoes, and $\varepsilon > 0$ be a given target amount of female mosquitoes. We will investigate the issue above for both Systems $(\mathcal{S}_1)$ and $(\mathcal{S}_2)$. We model the optimal control problems for the sterile mosquito release strategy as

$$\boxed{\inf_{u \in \mathcal{U}_{T,\overline{U},\varepsilon}^{(\mathcal{S}_1)}} J(u)}, \tag{$\mathcal{P}_{T,\overline{U},\varepsilon}^{(\mathcal{S}_1)}$}$$

and

$$\boxed{\inf_{u\in\mathcal{U}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon}} J(u)}, \tag{$\mathcal{P}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon}$}$$

where the functional J stands for the total number of released mosquitoes during the time T, namely

$$J(u) := \int_0^T u(t)dt.$$

For a given $\varepsilon > 0$, we introduce the sets of admissible controls for Systems ($\mathcal{S}_1$) and ($\mathcal{S}_2$), respectively denoted $\mathcal{U}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon}$ and $\mathcal{U}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon}$ and defined by

$$\mathcal{U}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon} := \Big\{u \in L^\infty(0,T) : 0 \leqslant u \leqslant \overline{U} \text{ a.e.}, F(T) \leqslant \varepsilon \\ \text{with } F \text{ solution of System } (\mathcal{S}_1)\Big\} \tag{5}$$

and

$$\mathcal{U}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon} := \Big\{u \in L^\infty(0,T) : 0 \leqslant u \leqslant \overline{U} \text{ a.e.}, F(T) \leqslant \varepsilon \\ \text{with } F \text{ solution of System } (\mathcal{S}_2)\Big\}.$$

Remark 3.1 (Exact null-controllability)**.** *It is notable that there does not exist any control $u \in L^\infty(0,T;\mathbb{R}_+)$ such that the corresponding solution F to System ($\mathcal{S}_1$) (resp. System ($\mathcal{S}_2$)) satisfies $F(T) = 0$ since one has $F(t) \geqslant F(0)e^{-\delta_F t}$ for every $t \in [0,T]$.*

Let us now state the main results of this article. In what follows, we will use the notation U^* to denote the positive number defined by (3), and the notations $J^{(\mathcal{S}_i)}_{T,\overline{U},\varepsilon}$, $i = 1,2$ to denote the optimal values for Problems ($\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon}$) and ($\mathcal{P}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon}$), namely

$$J^{(\mathcal{S}_i)}_{T,\overline{U},\varepsilon} := \inf_{u\in\mathcal{U}^{(\mathcal{S}_i)}_{T,\overline{U},\varepsilon}} J(u), \qquad i = 1,2. \tag{6}$$

Theorem 3.2. *Let us assume that Condition ($\mathcal{H}$) holds true. Let $\varepsilon \in (0,\overline{F})$ and $\overline{U} > U^*$. There exists a minimal time $\overline{T} > 0$ such that for all $T \geqslant \overline{T}$, the set $\mathcal{U}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon}$ is nonempty and Problem ($\mathcal{P}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon}$) has a solution u^*. One has $J^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon} \leqslant \overline{U}\,\overline{T}$ and the mappings $T \in [\overline{T},+\infty) \mapsto J^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon}$ and $\overline{U} \in (U^*,+\infty) \mapsto J^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon}$ are non-increasing. Furthermore, there exists $t_1 \in (0,T)$ such that*

$$u^* = 0 \quad \text{ on } (t_1,T),$$

one has $F(T) = \varepsilon$ and $F \in [\varepsilon,\overline{F}]$ on $(0,T)$.

It is interesting to note that there is no need to release sterile mosquitoes at the end of the experiment. For the simplified system ($\mathcal{S}_1$), we can obtain a much more precise characterization of optimal controls, which is the purpose of the next result. As a preliminary remark, recall that the function f is defined by (1).

Theorem 3.3. *Let us assume that Condition ($\mathcal{H}$) holds true and that $2\delta_s > \delta_F$. Let $\varepsilon \in (0,\overline{F})$, $\overline{U} > U^*$. There exists a minimal time $\overline{T} > 0$ such that for all $T \geqslant \overline{T}$, the set $\mathcal{U}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon}$ is nonempty and Problem ($\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon}$) has a solution u^*, characterized as follows:*

(i) *Optimal value: one has* $J^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon} \leqslant \overline{U}\,\overline{T}$ *and the mappings* $T \in [\overline{T},+\infty) \mapsto J^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon}$ *and* $\overline{U} \in (U^*,+\infty) \mapsto J^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon}$ *are non-increasing.*

(ii) *Optimal control: let* $T > \overline{T}$*. If* $\overline{U} > U^*$ *is large enough, there exists* $(t_0,t_1) \in [0,T]^2$ *with* $t_0 \leqslant t_1$ *such that*

(a) $u^* = 0$ *on* $(0,t_0)$ *or* $u^* = \overline{U}$ *on* $(0,t_0)$*;*

(b) $u^* \in (0,\overline{U})$ *on* (t_0,t_1) *with*

$$u^*(t) = \left(\frac{\partial^2 f}{\partial M_s^2}\right)^{-1} \left(\frac{\partial f}{\partial M_s}\frac{\partial f}{\partial F} + \delta_s M_s \frac{\partial^2 f}{\partial M_s^2} - f\frac{\partial^2 f}{\partial M_s \partial F}\right)\Bigg|_{(F(t),M_s(t))}; \tag{7}$$

(c) $u^* = 0$ *on* (t_1,T).

Furthermore, there exists $T^*(\overline{U})$ *(simply denoted* T^* *when no confusion is possible) such that* $T^* > \overline{T}$ *and that if* $T > T^*$*, then the optimizer* u^* *is unique, with* $0 < t_0 < t_1 < T$*, and such that* $u^* = 0$ *on* $(0,t_0)$*. Finally, the mapping* $T \in (T^*,+\infty) \mapsto J^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon}$ *is constant.*

(iii) *Optimal trajectory: we extend the domain of definition of* F *to* $\mathbb{R}_+$ *by setting* $u(\cdot) = 0$ *on* $(T,+\infty)$*. One has* $F(T) = \varepsilon$*,* $F' \leqslant 0$ *on* $(0,T)$ *and* $F' \geqslant 0$ *on* $(T,+\infty)$*. In particular,* F *has a unique local minimum at* T *which is moreover global, equal to* ε *and* $F'(T) = 0$.

Let us comment on this result. Under the assumptions of Theorem 3.3 (in particular that T and $\overline{U}$ are large enough), the infinite dimensional control problem ($\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon}$) can be reduced to a two dimensional one: more precisely, one only needs to determine the two parameters t_0 and t_1.

As expected, the horizon of time T fixed for the control to reach a desired number of adult female mosquitoes has a strong influence on the form of the optimal control. The longer T is, the smaller is the number of sterilized males needed. However, there is a maximal time T^* above which the number of released mosquitoes needed to reach the desired state is stationary with respect to the horizon of time.

Remark 3.4. *(Stabilization and feedback) Note that our results can be interpreted in terms of stabilization: indeed, it follows from Theorem 3.3, and in particular from the description of the optimal trajectory that, under the assumptions of this theorem, the differential system*

$$\begin{cases} \dfrac{d}{dt}\begin{pmatrix} F \\ M_s \end{pmatrix} = \begin{pmatrix} f(F,M_s) \\ u - \delta_s M_s \end{pmatrix} & \text{in } [0,+\infty) \\ \text{with } u = \dfrac{\frac{\partial f}{\partial M_s}(F,M_s)\frac{\partial f}{\partial F}(F,M_s) + \frac{\partial^2 f}{\partial M_s^2}(F,M_s)\delta_s M_s - \frac{\partial^2 f}{\partial M_s \partial F}(F,M_s) f(F,M_s)}{\frac{\partial^2 f}{\partial M_s^2}(F,M_s)} & , \end{cases}$$

complemented with the initial conditions $F(0) = \overline{F}$ *and* $M_s(0) = 0$*, satisfies*

$$\lim_{t\to+\infty} F(t) = 0,$$

in other words, the control function $u(\cdot)$ *above defines a feedback control stabilizing System* ($\mathcal{S}_1$).

3.2 A dedicated algorithm

In this section, we introduce an elementary algorithm for computing the (unique) solution to Problem $(\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$. The chosen approach mainly rests upon the precise description of the optimizer (whenever T and $\overline{U}$ are large enough) provided in Theorem 3.3.

Let us describe our approach. With the notations of Theorem 3.3, we assume that $T > T^*$ so that the optimal control is unique. We will take advantage of its particular form, more precisely that there exist $0 < t_0 < t_1 < T$ such that $u^* = 0$ on $(0, t_0)$, u^* is given by (7) on (t_0, t_1), $u^* = 0$ on (t_1, T) and $F(T) = \varepsilon$.

To this aim, let us introduce for $\tau_1 \in [0, T]$ the auxiliary Cauchy system

$$\left\{ \begin{array}{l} \dfrac{d}{dt}\begin{pmatrix} F \\ M_s \end{pmatrix} = \begin{pmatrix} f(F, M_s) \\ u_{\tau_1} - \delta_s M_s \end{pmatrix} \quad \text{in } [0, +\infty) \\ \text{with } u_{\tau_1} = \dfrac{\frac{\partial f}{\partial M_s}(F,M_s)\frac{\partial f}{\partial F}(F,M_s) + \frac{\partial^2 f}{\partial M_s^2}(F,M_s)\delta_s M_s - \frac{\partial^2 f}{\partial M_s \partial F}(F,M_s) f(F,M_s)}{\frac{\partial^2 f}{\partial M_s^2}(F,M_s)} \mathbb{1}_{(0,\tau_1)}, \end{array} \right. \tag{8}$$

complemented with the initial conditions $F(0) = \overline{F}$ and $M_s(0) = 0$, whose solution, associated to the control function u_{τ_1}, will from now on be denoted $(F^{\tau_1}, M_s^{\tau_1})$. Let $T > T^*$.

The uniqueness property for Problem $(\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$ allows us to get the following relations between the solution to Problem $(\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$ and the control u_{τ_1}.

Property 3.5. *Under the assumptions and with the notations of Theorem 3.3, let $\overline{U} > U^*$ (in particular we consider $T \geqslant T^*(\overline{U})$ so that the conclusion (ii) holds true). There exists $\varepsilon_0 > 0$ such that, if $\varepsilon \in (0, \varepsilon_0)$, then*

(i) for every $\tau_1 \in [0, T)$, there exists a unique $\tau_2(\tau_1) \in (\tau_1, +\infty)$ such that F^{τ_1} is first strictly decreasing on $(0, \tau_2(\tau_1))$, and then strictly increasing on $(\tau_2(\tau_1), +\infty)$.

(ii) the value function

$$\psi : \tau_1 \in [0, T) \mapsto \min_{t \in [0,\infty]} F^{\tau_1}(t) = F^{\tau_1}(\tau_2(\tau_1))$$

is decreasing.

(iii) the optimal control u^* solving Problem $(\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$ satisfies

$$u^*(t) = \left\{ \begin{array}{ll} 0 & \text{if } t \in (0, T - \tau_2(\tau_1)), \\ u_{\tau_1}(t - T + \tau_2(\tau_1)) & \text{otherwise} \end{array} \right. \tag{9}$$

a.e. on $(0, T)$, where τ_1 denotes the unique solution on $[0, T)$ to the equation $\psi(\tau_1) = \varepsilon$.

We now construct an efficient algorithm based on this result to solve Problem $(\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$, involving a bisection type method. The resulting Algorithm is described in Figure 2.

3.3 Numerical simulations

In order to illustrate our main results (Theorems 3.2 and 3.3), we provide some numerical simulations of the optimal control problems $(\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$ and $(\mathcal{P}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon})$. The codes are available on `github`[2].

[2] github.com/michelduprez/Optimal-control-strategies-for-the-sterile-mosquitoes-technique.git

Figure 2: Algorithm by dichotomy to solve Problem $(\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$

Initialization:

Let $n \in \mathbb{N}^*$, $\tau_{1,\min} = 0$ and $\tau_{1,\max} = T$

While $i \leq n$ **do**

1: $\tau_{1,\text{test}} = (\tau_{1,\min} + \tau_{1,\max})/2$
2: Solve (8) on $(0,T)$ for $\tau_1 = \tau_{1,\text{test}}$ and let u^{τ_1} be the function given by (8).
3: **if** $\min_{(0,T)} F^{\tau_1} < \varepsilon$ **then**
$\tau_{1,\max} = \tau_{1,\text{test}}$
4: **else**
$\tau_{1,\min} = \tau_{1,\text{test}}$

End:

Let $\tau_1^n = (\tau_{1,\min} + \tau_{1,\max})/2$, $t_1^n = \text{argmin}_{(0,T)} F^{\tau_1^n}$ and

$$u_n(t) = \begin{cases} 0 & \text{if } t \in (0, T - t_1^n), \\ u(t - T + t_1^n) & \text{otherwise,} \end{cases}$$

for each $t \in (0,T)$.

We use the parameter values provided in Table 1, that come from [29, Table 1-3]. As in [29], in order to get results relevant for an island of 74 ha (hectares) with an estimated male population of about 69 ha^{-1}, the total number of males at the beginning of the experiment is taken equal to $M^* = 69 \times 74 = 5106$. Regarding System $(\mathcal{S}_2)$, we assume that $M_s = 0$ at the equilibrium, and we then deduce that

$$\overline{E} = \frac{\delta_M}{(1-\nu)\nu_E}\overline{M}, \qquad \overline{F} = \frac{\nu\nu_E}{\delta_F}\overline{E} = \frac{\nu\delta_M}{\delta_F(1-\nu)}\overline{M}$$

and we thus evaluate the environmental capacity for eggs as

$$K = \left(1 - \frac{\delta_F(\nu_E + \delta_E)}{\beta_E \nu \nu_E}\right)^{-1} \overline{E}.$$

Regarding System $(\mathcal{S}_1)$, we consider the same initial quantity of females $\overline{F}$ and get that the environmental capacity K has the same expression as above.

To compute the solutions to Problem $(\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$ and $(\mathcal{P}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon})$, we use the opensource optimization routine GEKKO (see [8]) which solves the optimization problem thanks to the APOPT (Advanced Process OPTimizer) library, a software package for solving large-scale optimization problems (see [21]). We will also compare the numerical solution obtained when solving Problem $(\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$ with the one obtained by Formula (9), by applying our algorithm (see Figure 2).

Figures 3, 4 and 5 gather the solutions of the optimal control problems $(\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$, $(\mathcal{P}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon})$ and the function given by (9) for $T = 60, 150, 200$, $\overline{U} = 5000$, $\nu_E = 0.05$ and $\varepsilon = \overline{F}/4$.

The time interval $(0,T)$ is discretized with 300 points. We do not give the solution to (9) for $T = 60$ in Figure 3, since Algorithm 2 cannot be applied in this situation (indeed, T is not large enough and the form of the corresponding solution is not convenient to apply a bisection procedure).

For these parameter values, the bound U^* given in (3) is approximately equal to 9620. We remark that this bound is not optimal since, as we can see in Figure 3-5, the optimal control problems admits a solution for $\overline{U} = 5000$.

These simulations allow us to recover the theoretical results of Section 3. Indeed, the optimal control has the structure described in Theorem 3.3: at the beginning of the time interval, it vanishes (Figure 3) or it is equal to $\overline{U}$ (Figures 4-5), then we observe a singular part and finally a time interval for which the control vanishes (Figures 3-5).

Moreover, it is interesting to observe that, even if we do not have a proof of this fact, the optimal controls for the different problems look very close. This is confirmed by Table 2, where the quantity of sterile insects is computed for different values of ν_E. We remark also that ν_E has not a big influence on the total number of released mosquitoes.

We can observe that the length of the "waiting time" $T - \tau_2(\tau_1)$ for which $u^* = 0$ is not impacted by the total length of the time interval T. Indeed, this "waiting time" can be biologically and intuitively interpreted by the fact that the sterilized males continue to act after the last release.

Regarding Problem $(\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$, since for T large enough, the optimal control vanishes on an interval at the beginning of the experiment, the system stays at the equilibrium on this interval. Hence, with the notations of Theorem 3.3, the optimal time T_{opt} to control the system with a singular part is equal to $T - t_0$. As illustrated in Table 3, this optimal time T_{opt} is not very sensitive to the choice of optimal control problem, namely $(\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$, $(\mathcal{P}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon})$ and (9) and also of the value of ν_E.

In Table 4, we also provide, for different discretizations of the time interval, the computation time for solving Problems $(\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$ and $(\mathcal{P}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon})$ with the open-source optimization routine GEKKO (default parameters) and the one for solving (9) thanks to Algorithm 2 ($n = 50$ iterations). Here, we have used an `Intel CORE i5 8th Gen.` As expected, one can notice that Algorithm 2 yields a much faster resolution than the other approaches.

We finally give in Figure 6 the solution of the optimal control problem $(\mathcal{P}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon})$ for a small $\varepsilon = \overline{F}/1e4$ and a large final time T thanks to our numerical algorithm (see Figure 2) (the numerical resolution thanks to Gekko does not converge). We observe that in this case the control is increasing then decreasing. In other words, we act with large releases at the beginning of the time interval and then with small releases. Even if we do not take into account the Allee effect, we can remark that we do not need to release a lot of sterilized males at the end of the time interval.

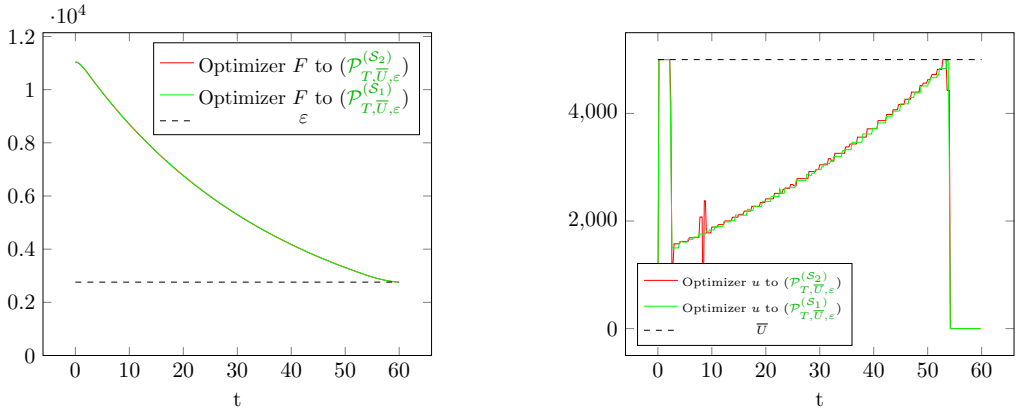

Figure 3: Solution of the optimal control problems $(\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$ and $(\mathcal{P}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon})$ with $T = 60$, $\overline{U} = 5000$, $\nu_E = 0.05$, and $\varepsilon = \overline{F}/4$.

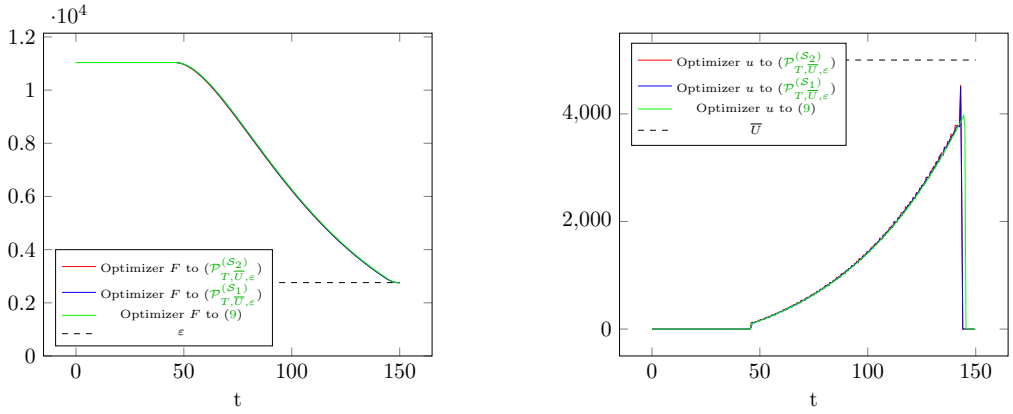

Figure 4: Solution of the optimal control problems $(\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$, $(\mathcal{P}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon})$ and (9) with $T = 150$, $\overline{U} = 5000$, $\nu_E = 0.05$, and $\varepsilon = \overline{F}/4$.

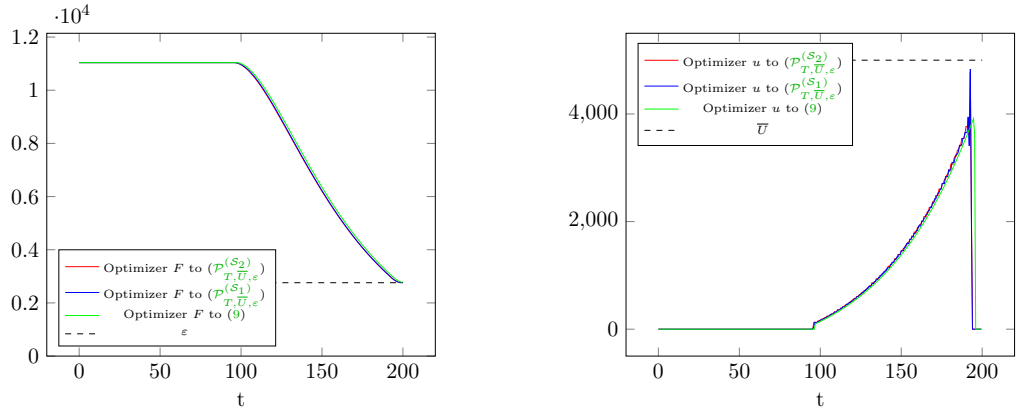

Figure 5: Solution of the optimal control problems $(\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$, $(\mathcal{P}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon})$ and (9) with $T = 200$, $\overline{U} = 5000$, $\nu_E = 0.05$, and $\varepsilon = \overline{F}/4$.

ν_E	0.005	0.01	0.02	0.03
$J(u^*)$ for $(\mathcal{P}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon})$	1.21e5	1.34e5	1.42e5	1.45e5
$J(u^*)$ for $(\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$	1.15e5	1.30e5	1.40e5	1.43e5
$J(u^*)$ for (9)	1.15e5	1.30e5	1.40e5	1.43e5

ν_E	0.05	0.1	0.15	0.25
$J(u^*)$ for $(\mathcal{P}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon})$	1.47e5	1.49e5	1.50e5	1.50e5
$J(u^*)$ for $(\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$	1.46e5	1.49e5	1.49e5	1.50e5
$J(u^*)$ for (9)	1.46e5	1.48e5	1.49e5	1.50e5

Table 2: $J(u^*)$ for $(\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$, $(\mathcal{P}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon})$ and (9) with respect to ν_E for $\varepsilon = \overline{F}/4$ with $\overline{F} = 11037$.

ν_E	0.005	0.01	0.02	0.03	0.05	0.1	0.15	0.25
T_{opt} for $(\mathcal{P}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon})$	109	107	105	105	104	104	104	104
T_{opt} for $(\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$	106	105	104	104	104	104	104	104
T_{opt} for (9)	106	105	104	103	103	103	103	103

Table 3: Optimal time T_{opt} of control for $(\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$, $(\mathcal{P}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon})$ and (9) with respect to ν_E for $\varepsilon = \overline{F}/4$ with $\overline{F} = 11037$.

Time discr.	50	100	200	400	800	1600	3200
C. t. for $(\mathcal{P}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon})$	2.64	8.11	2.56e1	1.86e2	3.83e4	X	X
C. t. for $(\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$	2.05	6.00	2.94e1	1.13e2	7.37e2	3.53e4	X
C. t. for (9)	3.58e-1	6.74e-1	1.44	2.73	5.45	11.7	2.24e1

Table 4: Time of computation (sec.) for $(\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$, $(\mathcal{P}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon})$ and (9) with respect to the time discretization for $\nu_E = 0.05$, $\varepsilon = \overline{F}/4$ with $\overline{F} = 11037$.

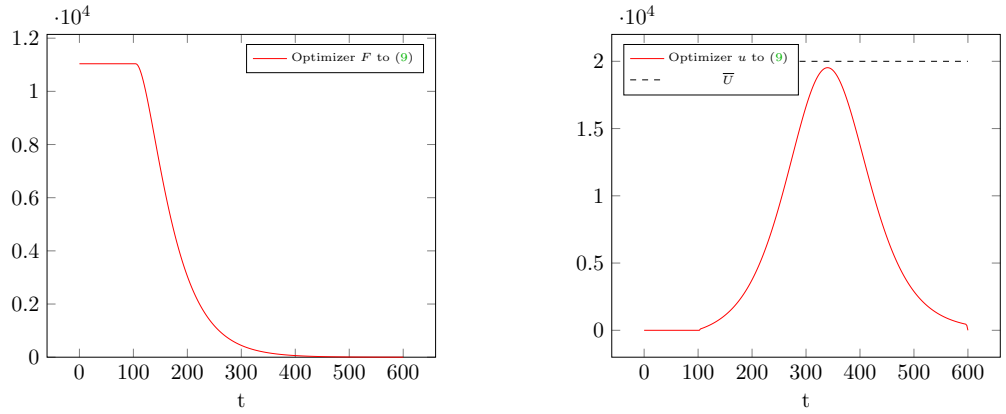

Figure 6: Solution of the optimal control problems (9) with $T = 600$, $\overline{U} = 20000$, $\nu_E = 0.05$, and $\varepsilon = \overline{F}/1e4$.

4 Proofs of the main results

4.1 Proof of Theorem 3.2

In the whole proof, we will denote by $(f_E(E,F),\ f_M(E,M),\ f_F(E,M,F,M_s))^\top$ the right-hand side of the differential subsystem of System $(\mathcal{S}_2)$ satisfied by $(E,M,F)^\top$, namely

$$f_E(E,F) = \beta_E F\left(1 - \frac{E}{K}\right) - (\nu_E + \delta_E)E,$$

$$f_M(E,M) = (1-\nu)\nu_E E - \delta_M M$$

and

$$f_F(E,M,F,M_s) = \nu\nu_E E\frac{M}{M+\gamma_s M_s} - \delta_F F.$$

We first point out that system $(\mathcal{S}_2)$ enjoys a monotonicity property with respect to the control u.

Lemma 4.1. *Let $u_1, u_2 \in L^\infty(0,T;\mathbb{R}_+)$ be such that $u_1 \geqslant u_2$. Let us assume that (4) holds.*

Then, the associated solutions (E_1,M_1,F_1,M_{s1}), (E_2,M_2,F_2,M_{s2}) to System $(\mathcal{S}_2)$ respectively associated to u_1 and u_2 satisfy $(E_1,M_1,F_1) \leqslant (E_2,M_2,F_2)$, the inequality being understood component by component.

Proof. According to Proposition 2.1, we have $(E,M,F) \in [0,\overline{E}]\times[0,\overline{M}]\times[0,\overline{F}]$. Noting that

$$M_{s_i}(t) = \int_0^t u_i(s)e^{\delta_s(s-t)}\,ds, \quad i=1,2,$$

it follows that $M_{s1} \geqslant M_{s2}$. Hence, the monotonicity property follows since one has

$$\frac{\partial f_E}{\partial F} \geqslant 0, \quad \frac{\partial f_M}{\partial E} \geqslant 0, \quad \frac{\partial f_F}{\partial E} \geqslant 0, \quad \frac{\partial f_F}{\partial M} \geqslant 0, \quad \frac{\partial f_F}{\partial M_s} \leqslant 0,$$

and therefore the so-called Kamke-Müller conditions (see e.g. [22]) hold true. □

Let us investigate the existence of an optimal control for Problem $(\mathcal{P}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon})$.

Lemma 4.2. *Let $\varepsilon \in (0,\overline{F})$ and $\overline{U} > U^*$. There exists $\overline{T}(\overline{U}) > 0$ such that for all $T \geqslant \overline{T}(\overline{U})$, the set $\mathcal{U}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon}$ is nonempty and for such a choice of T, Problem $(\mathcal{P}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon})$ has a solution u^*.*

Moreover, one has $J^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon} \leqslant \overline{U}\,\overline{T}$ (where $J^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon}$ is defined by (6)) and $J^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon}$ is non-increasing with respect to $T \geqslant \overline{T}$ and $\overline{U} > U^*$.

Proof. According to Proposition 2.1, if $u = \overline{U}$, then $F(t) \to 0$ as t goes to $+\infty$. Hence, for any $\varepsilon \in (0,\overline{F})$ and since F is Lipschitz-continuous, there exists $\overline{T} > 0$ such that $F(\overline{T}) \leqslant \varepsilon$ meaning that $T \geqslant \overline{T}$, $u = \overline{U}\mathbb{1}_{[T-\overline{T},T]} \in \mathcal{U}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon}$. The set $\mathcal{U}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon}$ is thus nonempty.

Let us now investigate the existence property. For the sake of clarity, we temporarily denote by (E^u,M^u,F^u,M^u_s) the solution of System $(\mathcal{S}_2)$ associated to u. Let $(u_n)_{n\in\mathbb{N}}$ be a minimizing sequence. According to the Banach-Alaoglu Bourbaki theorem (see e.g. [27]), the set $\mathcal{U}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon}$ is compact for the weak-$*$ topology of $L^\infty(0,T)$ and, up to a subsequence, $(u_n)_{n\in\mathbb{N}}$ converges to $u \in L^\infty(0,T,[0,\overline{U}])$. The functional J is obviously continuous for the weak-$*$ topology of $L^\infty(0,T)$, so that it only remains to prove that $u \in \mathcal{U}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon}$, in other words that $F^u(T) \leqslant \varepsilon$. First, one has

$$M^{u_n}_s(t) = \int_0^t e^{-\delta_s(t-s)}u_n(s)\,ds$$

for all $t \geqslant 0$ and it thus follows that, $(M^{u_n}_s)_{n\in\mathbb{N}}$ is bounded in $W^{1,+\infty}(0,T)$. Thus, it converges up to a subsequence to M^u_s strongly in $C^0([0,T])$ according to the Ascoli theorem (see e.g. [27]). By using the same reasoning as above, the sequence $(E^{u_n},M^{u_n},F^{u_n})$ is non-negative, uniformly bounded by above and therefore, the right-hand side of the first three equations of $(\mathcal{S}_2)$ is bounded in $W^{1,+\infty}(0,T)$. Using to the Ascoli theorem, we infer that $(F^{u_n})_{n\in\mathbb{N}}$ converges to F^u in $C^0([0,T])$. The desired conclusion follows by passing to the limit in the inequality $F^{u_n}(T) \leqslant \varepsilon$.

Finally, the monotonicity of $J^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon}$ with respect to $\overline{U}$ comes from the monotonicity of the admissible control set $\mathcal{U}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon}$ with respect to $\overline{U}$ and T for the inclusion. More precisely, if $\overline{U}_1 \leqslant \overline{U}_2$ then $\mathcal{U}^{(\mathcal{S}_2)}_{T,\overline{U}_1,\varepsilon} \subset \mathcal{U}^{(\mathcal{S}_2)}_{T,\overline{U}_2,\varepsilon}$ according to Lemma 4.1. Moreover, if $T_1 \leqslant T_2$, let u_1^* be a solution of Problem $(\mathcal{P}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon})$. Then, $u_2 = u_1^*(\cdot - T_2 + T_1)\mathbb{1}_{(T_2-T_1,T_2)} \in \mathcal{U}_{T_2,\overline{U},\varepsilon}$ (since (E, M, F) is stationary on $[0, T_2 - T_1)$), and is such that $J(u_2) = J^{(\mathcal{S}_2)}_{T_1,\overline{U},\varepsilon}$. Hence, we get by minimality that $J^{(\mathcal{S}_2)}_{T_2,\overline{U},\varepsilon} \leqslant J^{(\mathcal{S}_2)}_{T_1,\overline{U},\varepsilon}$.

Finally, since $\overline{u} = \overline{U}\mathbb{1}_{[T-\overline{T},T]}$ belongs to $\mathcal{U}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon}$ for any $\overline{U} \geqslant U^*$ and $T \geqslant \overline{T}(\overline{U})$, one has $J^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon} \leqslant J(\overline{u}) = \overline{U}\,\overline{T}$. □

In the following result, one shows that the state constraint is reached by any optimal control.

Lemma 4.3. *Let $\overline{U} > U^*$ and $T \geqslant \overline{T}(\overline{U})$. Let u^* solving Problem $(\mathcal{P}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon})$ and (E, M, F, M_s) be the corresponding solution to System $(\mathcal{S}_2)$. Then, one has necessarily $F(T) = \varepsilon$ and $F(\cdot) \in (\varepsilon, \overline{F}]$ on $[0, T)$.*

Proof. Let u^* be a solution to Problem $(\mathcal{P}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon})$. Since $\overline{F}$ is a stationary solution and $F(0) = \overline{F}$, one has $F \leqslant \overline{F}$ on $(0, T)$ by using Proposition 2.3. It remains to prove that F is bounded below by ε. Let us assume by contradiction that the corresponding solution (E, M, F, M_s) to System $(\mathcal{S}_2)$ satisfies $F(T) < \varepsilon$. Let $u_\eta := (1 - \eta)u^*$ with $\eta \in (0, 1)$ and denote by $(E_\eta, M_\eta, F_\eta, M_{s\eta})$ the corresponding solution to System $(\mathcal{S}_2)$. Then, by mimicking the arguments of the proof of Lemma 4.2, one easily gets that the mapping $u \in \mathcal{U}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon} \mapsto (E, M, F, M_s) \in C^0([0, T])^4$ is continuous for the weak-$*$ topology of L^∞ and it follows that $F_\eta(T) = F(T) + \mathrm{O}(\eta)$ so that $F_\eta(T) < \varepsilon$ whenever η is small enough. Since $\int_0^T u_\eta(t)\,dt < \int_0^T u^*(t)\,dt$, this contradicts the minimality of u^*.

Finally, we claim that $F(\cdot) > \varepsilon$ on $[0, T)$. In the converse case, there exists $T' < T$ such that $F(T') = \varepsilon$. One sees easily that the control $u_{T'}$ defined by $u_{T'} = u^*(\cdot - T + T')\mathbb{1}_{[T-T',T]}$ belongs to $\mathcal{U}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon}$ and that $J(u_{T'}) < J(u^*)$, whence a contradiction. □

Let us now state the necessary optimality conditions for Problem $(\mathcal{P}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon})$. To this aim, let us introduce (P, Q, R, S) as the solution to the backward adjoint system

$$\left\{\begin{array}{l} -\dfrac{d}{dt}\begin{pmatrix} P \\ Q \\ R \\ S \end{pmatrix} = \begin{pmatrix} \frac{\partial f_E}{\partial E}(E,F) & \frac{\partial f_M}{\partial E}(E,M) & \frac{\partial f_F}{\partial E}(E,M,F,M_s) & 0 \\ 0 & \frac{\partial f_M}{\partial M}(E,M) & \frac{\partial f_F}{\partial M}(E,M,F,M_s) & 0 \\ \frac{\partial f_E}{\partial F}(E,F) & 0 & \frac{\partial f_F}{\partial F}(E,M,F,M_s) & 0 \\ 0 & 0 & \frac{\partial f_F}{\partial M_s}(E,M,F,M_s) & -\delta_s \end{pmatrix}\begin{pmatrix} P \\ Q \\ R \\ S \end{pmatrix}, \\ P(T) = 0, Q(T) = 0, R(T) = 1,\ S(T) = 0. \end{array}\right. \tag{10}$$

Let us first determine the differential of $F(T)$ with respect to the control u.

Lemma 4.4. *Let $u \in \mathcal{U}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon}$ and introduce the functional G defined by $G(u) = F(T)$, where (E, M, F, M_s) denotes the unique solution of System $(\mathcal{S}_2)$ associated to u. Then, G is differentiable in the sense of Fréchet and for every admissible perturbation[3] h, the Gâteaux-derivative of G*

[3]More precisely, we call "admissible perturbation" any element of the tangent cone $\mathcal{T}_{u,\mathcal{U}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon}}$ to the set $\mathcal{U}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon}$ at u. Recall that the cone $\mathcal{T}_{u,\mathcal{U}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon}}$ is the set of functions $h \in L^\infty(0, T)$ such that, for any sequence of positive real numbers ε_n decreasing to 0, there exists a sequence of functions $h_n \in L^\infty(0, T)$ converging to h as $n \to +\infty$, and $u + \varepsilon_n h_n \in \mathcal{U}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon}$ for every $n \in \mathbb{N}$ (see e.g. [13]).

at u in the direction h is

$$DG(u)\cdot h=\int_0^T h(t)S(t)dt,$$

where (P,Q,R,S) solves System (10).

Proof. The Fréchet-differentiability of G is standard and follows from the differentiability of the mapping $u\in\mathcal{U}_{T,\overline{U},\varepsilon}^{(\mathcal{S}_2)}\mapsto F$, where (E,M,F,M_s) denotes the unique solution of System $(\mathcal{S}_2)$, itself deriving from a standard application of the implicit function theorem combined with variational arguments.

Moreover, the Fréchet-derivative $(\dot{E},\dot{M},\dot{F},\dot{M}_s)$ of (E,M,F,M_s) at u in the direction h solves the linearized problem

$$\begin{cases}\dfrac{d}{dt}\begin{pmatrix}\dot{E}\\ \dot{M}\\ \dot{F}\\ \dot{M}_s\end{pmatrix}=A\begin{pmatrix}\dot{E}\\ \dot{M}\\ \dot{F}\\ \dot{M}_s\end{pmatrix}+\begin{pmatrix}0\\0\\0\\h\end{pmatrix},\\ \dot{E}(0)=0,\ \dot{M}(0)=0,\ \dot{F}(0)=0,\ \dot{M}_s(0)=0,\end{cases}$$

where the Jacobian matrix A reads

$$A=\begin{pmatrix}\frac{\partial f_E}{\partial E}(E,F) & 0 & \frac{\partial f_E}{\partial F}(E,F) & 0\\ \frac{\partial f_M}{\partial E}(E,M) & \frac{\partial f_M}{\partial M}(E,M) & 0 & 0\\ \frac{\partial f_F}{\partial E}(E,M,F,M_s) & \frac{\partial f_F}{\partial M}(E,M,F,M_s) & \frac{\partial f_F}{\partial F}(E,M,F,M_s) & \frac{\partial f_F}{\partial M_s}(E,M,F,M_s)\\ 0 & 0 & 0 & -\delta_s\end{pmatrix}.$$

By integration by parts, it follows that

$$\begin{aligned}\int_0^T\left\langle\begin{pmatrix}0\\0\\0\\h\end{pmatrix},\begin{pmatrix}P\\Q\\R\\S\end{pmatrix}\right\rangle &= \int_0^T\left\langle\frac{d}{dt}\begin{pmatrix}\dot{E}\\ \dot{M}\\ \dot{F}\\ \dot{M}_s\end{pmatrix},\begin{pmatrix}P\\Q\\R\\S\end{pmatrix}\right\rangle-\int_0^T\left\langle A\begin{pmatrix}\dot{E}\\ \dot{M}\\ \dot{F}\\ \dot{M}_s\end{pmatrix},\begin{pmatrix}P\\Q\\R\\S\end{pmatrix}\right\rangle\\ &= -\int_0^T\left\langle\begin{pmatrix}\dot{E}\\ \dot{M}\\ \dot{F}\\ \dot{M}_s\end{pmatrix},\frac{d}{dt}\begin{pmatrix}P\\Q\\R\\S\end{pmatrix}\right\rangle+\left[\left\langle\begin{pmatrix}\dot{E}\\ \dot{M}\\ \dot{F}\\ \dot{M}_s\end{pmatrix},\begin{pmatrix}P\\Q\\R\\S\end{pmatrix}\right\rangle\right]_0^T\\ &\quad -\int_0^T\left\langle\begin{pmatrix}\dot{E}\\ \dot{M}\\ \dot{F}\\ \dot{M}_s\end{pmatrix},A^T\begin{pmatrix}P\\Q\\R\\S\end{pmatrix}\right\rangle=F(T),\end{aligned}$$

which leads to the desired result. □

Remark 4.5. *This result, as well as the next one, can also be obtained by using the so-called Pontryagin Maximum Principle (see for instance [24]). In particular, the final condition for the adjoint state is called transversality condition.*

The following Lemma leads to a characterization of any optimal control. We chose here to provide some quick explanations on the derivation of optimality conditions. It would have also been possible to use a shorter argument through the so-called Pontryagin Maximum Principle (PMP) and we would have obtained the same result.

Lemma 4.6. *Let $\overline{U} > U^*$ and $T \geqslant \overline{T}(\overline{U})$. Let u^* denote a solution to Problem $(\mathcal{P}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon})$. There exists $\lambda > 0$ such that*

$$\begin{cases} \text{a.e. on } \{u^* = 0\}, \text{ one has } 1 + \lambda S(t) \geqslant 0, \\ \text{a.e. on } \{0 < u^* < \overline{U}\}, \text{ one has } 1 + \lambda S(t) = 0, \\ \text{a.e. on } \{u^* = \overline{U}\}, \text{ one has } 1 + \lambda S(t) \leqslant 0. \end{cases}$$

Proof. Let us introduce the Lagrangian function $\mathcal{L}$ associated to problem $(\mathcal{P}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon})$, defined by

$$\mathcal{L} : (u, \lambda) \in \mathcal{U}_{T,\overline{U}} \times \mathbb{R} \mapsto J(u) - \lambda(F(T) - \varepsilon),$$

where $\mathcal{U}_{T,\overline{U}} := \{u \in L^\infty(0,T) : 0 \leqslant u(\cdot) \leqslant \overline{U}\}$.

By standard arguments, we get the existence of a Lagrange multiplier $\lambda \geqslant 0$ such that (u^*, λ) satisfies $D_u\mathcal{L}(u^*, \lambda) \cdot h \geq 0$ for every h belonging to the tangent cone of the set $\mathcal{U}^{(\mathcal{S}_2)}_{T,\overline{U}}$ at u^*. Moreover, according to Lemma 4.3, we have necessarily $F(T) = \varepsilon$.

Let t^* be a Lebesgue density-one point of $\{u^* = 0\}$. Let $(H_n)_{n\in\mathbb{N}}$ be a sequence of measurable subsets containing all t^* and such that H_n is included in $\{u^* = 0\}$. Let us consider $h = \mathbb{1}_{H_n}$ and notice that, by construction, $u^* + \eta h$ belongs to $\mathcal{U}^{(\mathcal{S}_2)}_{T,\overline{U}}$ whenever η is small enough. One has

$$\mathcal{L}(u^* + \eta h, \lambda) \geqslant \mathcal{L}(u^*, \lambda),$$

whenever η is small enough. Let us divide this inequality by η, and let η go to 0. By using Lemma 4.4, we obtain

$$\int_0^T h(t)dt + \lambda \int_0^T h(t)S(t)dt \geqslant 0,$$

which rewrites $|H_n| + \lambda \int_{H_n} S(t)dt \geqslant 0$. Dividing this inequality by $|H_n|$ and letting H_n shrink to $\{t^*\}$ as $n \to \infty$ shows that $1 + \lambda S(t) \geqslant 0$ on $\{u^* = 0\}$. This proves the first point of Lemma 4.6, according to the Lebesgue Density Theorem (see e.g. [26]). The proof of the third point is similar and consists in considering perturbations of the form $u^* - \eta h$, where h denotes a positive admissible perturbation of u^* supported in $\{u^* = \overline{U}\}$. Finally, the proof of the second point follows the same lines by considering bilateral perturbations of the form $u^* \pm \eta h$, where h denotes an admissible perturbation of u^* supported in $\{0 < u^* < \overline{U}\}$.

Let us now prove that $\lambda > 0$. We argue by contradiction, assuming that $\lambda = 0$. Then, the switching function $1 + \lambda S$ is necessarily constant, equal to 1, and we have therefore $u^* = 0$ in $[0, T]$, which leads to a contradiction since the optimal trajectory has to satisfy $F(T) = \varepsilon$. □

Let us prove the remaining facts stated in Theorem 3.2.

Proof of Theorem 3.2. Let $\varepsilon \in (0, \overline{F})$. According to Lemma 4.2, there exists U^* such that for each $\overline{U} > U^*$, there exists $\overline{T}$ such for all $T > \overline{T}$, Problem $(\mathcal{P}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon})$ has a solution u^*. Let (E, M, F, M_s) be the optimal trajectory. According to Lemma 4.3, the constraint "$F(T) = \varepsilon$" is reached. Lemma 4.6 implies that on $\{S > -1/\lambda\}$, one has necessarily $u = 0$. Since S is continuous and $S(T) = 0$, it follows that there exists $t_1 \in (0, T)$ such that $u^* = 0$ on (t_1, T). □

4.2 Proof of Theorem 3.3

Let us first point out that System $(\mathcal{S}_1)$ is monotone with respect to the control u :

Lemma 4.7. *Let $u_1, u_2 \in L^\infty(0,T)$ such that $u_1 \geqslant u_2 \geqslant 0$ (resp. $u_1 > u_2 \geqslant 0$). Let us assume that (4) holds. Then, the corresponding solutions F_1, F_2 to System $(\mathcal{S}_1)$ satisfy $F_1 \leqslant F_2$ on $(0,T)$ (resp. $F_1 < F_2$ on $(0,T)$).*

Proof. This is an immediate consequence of the fact that $\frac{\partial f}{\partial M_s} \leqslant 0$ when $F \in (0, \overline{F}]$. □

Let us investigate the existence of an optimal control for Problem $(\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$.

Lemma 4.8. *Let $\varepsilon \in (0, \overline{F})$. For every $\overline{U} > U^*$, there exists $\overline{T}(\overline{U}) > 0$ such that for all $T \geqslant \overline{T}(\overline{U})$, the set $\mathcal{U}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon}$ is nonempty and Problem $(\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$ has a solution u^*.*

Moreover, one has $J^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon} \leqslant \overline{U}\,\overline{T}$ (where $J^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon}$ is defined by (6)) and $J^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon}$ is non-increasing with respect to $T \geqslant \overline{T}$ and to $\overline{U} > U^*$.

Proof. We first prove that $\mathcal{U}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon}$ is nonempty when $\overline{U} > U^*$ and T is large enough. If $u(\cdot) = \overline{U} > U^*$, we have seen in Proposition 2.3 that $F(t) \to 0$ as $t \to +\infty$. Thus, for any $\varepsilon \in (0, \overline{F})$, there exists $\overline{T} > 0$, such that $F(\overline{T}) = \varepsilon$. Hence, for $T \geqslant \overline{T}$, $u = \overline{U}\mathbb{1}_{[T-\overline{T},T]}$ belongs to $\mathcal{U}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon}$.

Proceeding as in the proof of Lemma 4.2, we obtain existence of a solution to Problem $(\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$ by considering a minimizing sequence and showing that it is in fact compact. Finally, the monotonicity and the bound on $J^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon}$ are obtained exactly as in the end of the proof of Lemma 4.2. □

By mimicking the proof of Lemma 4.3 for Problem $(\mathcal{P}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon})$, we can prove that the constraint $F(T) \leqslant \varepsilon$ is saturated.

Lemma 4.9. *Let $\overline{U} > U^*$, $T \geqslant \overline{T}(\overline{U})$ and u^* be a solution to Problem $(\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$. Let (F, M_s) be the associated optimal trajectory, solution to System $(\mathcal{S}_1)$. Then, one has $F(T) = \varepsilon$ and $F(\cdot) \in (\varepsilon, \overline{F}]$ on $[0,T)$.*

We can also prove that F is non-increasing on $(0,T)$.

Lemma 4.10. *Let $\overline{U} > U^*$, $T \geqslant \overline{T}(\overline{U})$ and u^* be a solution to Problem $(\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$. Let (F, M_s) be the associated optimal trajectory, solution to System $(\mathcal{S}_1)$. Then, for every $t \in (0,T)$, one has $F'(t) \leqslant 0$.*

Proof. Let us argue by contradiction, assuming that the conclusion is not true. Then, since F is C^1, there exist $0 < \theta_1 < \theta_2 < T$ such that $F' > 0$ on (θ_1, θ_2).

According to Lemma 4.7, F is non-increasing in a neighborhood of 0. Moreover, according to Lemma 4.9, F decreases to ε in a neighborhood of T. Consequently, it is not restrictive to also assume that $F'(\theta_1) = F'(\theta_2) = 0$, and $F' \leqslant 0$ on $(0, \theta_1)$

Notice from the expression of f given in (1) that $F' = f(F, M_s) \leqslant 0$ if, and only if,

$$\delta_M \gamma_s M_s \geqslant \phi(F) = (1-\nu)\beta_E \nu_E \frac{\nu\beta_E\nu_E - \delta_F\left(\frac{\beta_E F}{K} + \nu_E + \delta_E\right)}{\delta_F\left(\frac{\beta_E F}{K} + \nu_E + \delta_E\right)^2} F.$$

Let us show that there exist $0 < \tau_1 < \tau_2 < T$ such that $F(\tau_1) < F(\tau_2)$ and $M_s(\tau_1) \geqslant M_s(\tau_2)$. Indeed, there are two possibilities. Either there exists $\tau_2 \in (\theta_1, \theta_2)$ such that $M_s(\theta_1) = M_s(\tau_2)$, and then we take $\tau_1 = \theta_1$, or for any $t \in (\theta_1, \theta_2)$, $M_s(\theta_1) < M_s(t)$. In this latter case, we take $\tau_2 \in (\theta_1, \theta_2)$ such that $M_s(\theta_1) < M_s(\tau_2) < \phi(F(\tau_2))$ (which is always possible since $F' > 0$ on (θ_1, θ_2)). Then, since $F(0) = \overline{F} > F(\tau_2) > F(\theta_1)$ and F is continuous, there exists $\tilde{\tau} \in (0, \theta_1)$ such that $F(\tilde{\tau}) = F(\tau_2)$. Moreover, since $M_s \geqslant \phi(F)$ on $(0, \theta_1)$, we have $M_s(\tilde{\tau}) \geqslant \phi(F(\tilde{\tau})) = \phi(F(\tau_2)) > M_s(\tau_2) > M_s(\theta_1)$.

By continuity of M_s, there exists $\tau_1 \in (\tilde{\tau}, \theta_1)$ such that $M_s(\tau_1) = M_s(\tau_2)$. Since we have $F' \leqslant 0$ on $(0, \theta_1)$, we deduce that $F(\tau_1) \leqslant F(\tilde{\tau}) = F(\tau_2)$ and $F(\tau_2) \neq F(\tau_1)$ since $\phi(F(\tau_1)) \leqslant M_s(\tau_1) = M_s(\tau_2) < \phi(F(\tau_2))$.

Then, we take

$$u(t) = \begin{cases} 0, & \text{for } t \in (0, \tau_2 - \tau_1), \\ u^*(t - \tau_2 + \tau_1), & \text{for } t \in (\tau_2 - \tau_1, \tau_2), \\ u^*(t), & \text{for } t \in (\tau_2, T). \end{cases}$$

Using Lemma 4.13, we have $J(u) \leqslant J(u^*)$. Moreover, if we denote by (F^u, M_s^u) the solution with this control u, we have $F^u(\tau_2) = F(\tau_1) < F(\tau_2)$ and $M_s^u(\tau_2) = M_s(\tau_1) \geqslant M_s(\tau_2)$. Then, on (τ_2, T), we have $M_s^u \geqslant M_s$ and since $M_s \mapsto f(F, M_s)$ is non-increasing $(F^u)' = f(F^u, M_s^u) \leqslant f(F^u, M_s)$. By comparison with the solution to $F' = f(F, M_s)$ on (τ_2, T) and since $F^u(\tau_2) < F(\tau_2)$, we deduce that $F^u(T) < F(T) = \varepsilon$ which contradicts Lemma 4.9. □

Let us introduce (Q, R) as the solution of the forward adjoint system

$$\begin{cases} -\dfrac{d}{dt}\begin{pmatrix} Q \\ R \end{pmatrix} = \begin{pmatrix} \frac{\partial f}{\partial F}(F, M_s) & 0 \\ \frac{\partial f}{\partial M_s}(F, M_s) & -\delta_s \end{pmatrix}\begin{pmatrix} Q \\ R \end{pmatrix}, \\ Q(T) = 1, \ R(T) = 0. \end{cases} \tag{11}$$

Similarly to Problem $(\mathcal{P}^{(\mathcal{S}_2)}_{T,\overline{U},\varepsilon})$, any optimal control can be characterized by using the first order necessary optimality conditions, in terms of a switching function of the form $t \mapsto 1 + \lambda R(t)$ with $\lambda \geq 0$.

Lemma 4.11. *Let $\overline{U} > U^*$ and $T \geqslant \overline{T}(\overline{U})$. Consider u^* a solution to Problem $(\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$. Then there exists $\lambda > 0$ such that*

$$\begin{cases} \text{a.e. on } \{u^* = 0\}, \text{ one has } 1 + \lambda R(t) \geqslant 0, \\ \text{a.e. on } \{0 < u^* < \overline{U}\}, \text{ one has } 1 + \lambda R(t) = 0, \\ \text{a.e. on } \{u^* = \overline{U}\}, \text{ one has } 1 + \lambda R(t) \leqslant 0. \end{cases}$$

The proof is similar to the one of Lemma 4.6.

Lemma 4.12. *Let $\overline{U} > U^*$ and $T \geqslant \overline{T}(\overline{U})$. There exists $t_1 \in (0, T)$ such that $u^* = 0$ on (t_1, T).*

Proof. According to Lemma 4.11, on the set $\{R > -1/\lambda\}$, one has necessarily $u = 0$. We conclude using the fact that R is continuous and $R(T) = 0$. □

Lemma 4.13. *Let $\overline{U} > U^*$ and $T \geqslant \overline{T}(\overline{U})$. There exist positive constants C_0, C_1, $C_2 > 0$ that do not depend on $\overline{U}$ and T, such that*

$$|M_s| < C_0, \quad 0 < C_1 \leqslant Q \leqslant C_2.$$

Proof. First, one has for all $t \in [0, T]$,

$$M_s(t) = \int_0^t u^*(s) e^{\delta_s(s-t)} ds \leqslant \int_0^T u^*(t)\, dt \leqslant \overline{TU}.$$

From (11), Q solves

$$-\frac{dQ}{dt} = \frac{\partial f}{\partial F}(F, M_s)Q, \qquad Q(T) = 1.$$

Note that, on $[0, T]$, all quantities are bounded: one has $\varepsilon \leqslant F(\cdot) \leqslant F(0) = \overline{F}$ according to Lemma 4.10. Hence, we infer the existence of $\mu > 0$ such that $-\mu \leqslant \frac{\partial f}{\partial F}(F, M_s) \leqslant \mu$. Thus, integrating the inequality $-\mu Q \leqslant -\frac{dQ}{dt} \leqslant \mu Q$ with $Q(T) = 1$, the conclusion follows easily by using a Gronwall type argument (see e.g. [20]). □

Lemma 4.14. *Let $\overline{U} > U^*$ and $T \geqslant \overline{T}(\overline{U})$. If $0 < u^* < \overline{U}$ on a nonempty interval (s_0, s_1) then u^* satisfies (7) on (s_0, s_1).*

Proof. According to Lemma 4.11, one has $R' = 0$ on (s_0, s_1). Differentiating the equation satisfied by R in (11) yields

$$\left(\frac{\partial^2 f}{\partial M_s \partial F} F' + \frac{\partial^2 f}{\partial M_s^2} M_s'\right) Q + \frac{\partial f}{\partial M_s} Q' = 0.$$

We deduce that

$$\left(\frac{\partial^2 f}{\partial M_s \partial F} f + \frac{\partial^2 f}{\partial M_s^2}(u - \delta_s M_s)\right) Q = \frac{\partial f}{\partial M_s}\frac{\partial f}{\partial F} Q.$$

According to Lemma 4.13, one has $Q > 0$ on $[0, T]$. Hence u is given by

$$u = \left(\frac{\partial^2 f}{\partial M_s^2}\right)^{-1} \left(\frac{\partial f}{\partial M_s}\frac{\partial f}{\partial F} + \frac{\partial^2 f}{\partial M_s^2}\delta_s M_s - \frac{\partial^2 f}{\partial M_s \partial F} f\right),$$

where f is given by (1). We thus recover (7). □

Lemma 4.15. *Let $\overline{U} > U^*$ and $T \geqslant \overline{T}(\overline{U})$.*

(i) If $2\delta_s > \delta_F$, then, on each open interval of $\{R > -1/\lambda\}$, any local extremum for R is a minimum.

(ii) If $\overline{U}$ is large enough, then, on each open interval of $\{R < -1/\lambda\}$, any local extremum for R is a maximum.

Proof. (i) By differentiating the equation on R in (11), we get

$$-R'' = \left(\frac{\partial^2 f}{\partial M_s \partial F} f + \frac{\partial^2 f}{\partial M_s^2}(u^* - \delta_s M_s) - \frac{\partial f}{\partial M_s}\frac{\partial f}{\partial F}\right) Q - \delta_s R' \quad \text{on } (0, T). \tag{12}$$

Let I be an open interval of $\{R > -1/\lambda\}$ (whenever it exists). Let $\tau \in I$ be a local extremum for R, we have $R'(\tau) = 0$. Then, thanks to Lemma 4.11, we have $u^* = 0$ on I and hence

$$-R''(\tau) = \left(\frac{\partial^2 f}{\partial M_s \partial F} f - \frac{\partial^2 f}{\partial M_s^2}\delta_s M_s - \frac{\partial f}{\partial M_s}\frac{\partial f}{\partial F}\right)(\tau) Q(\tau).$$

We have seen in Lemma 4.13 that $Q > 0$. Then, the sign of $-R''(\tau)$ is the sign of $\mathfrak{U}$ given by

$$\mathfrak{U} = \frac{\partial^2 f}{\partial M_s \partial F} f - \frac{\partial^2 f}{\partial M_s^2}\delta_s M_s - \frac{\partial f}{\partial M_s}\frac{\partial f}{\partial F}.$$

Let us compute $\mathfrak{U}$. Notice that f is of the form

$$f(F, M_s) = \mu F^2 \Lambda - \delta_F F \qquad \text{where} \qquad \Lambda := \frac{1}{F^2 + aF + M_s(\alpha F^2 + \beta F + \gamma)} \tag{13}$$

with some positive constants $\mu, a, \alpha, \beta, \gamma$. Observe that

$$\frac{\partial \Lambda}{\partial M_s} = -\Lambda^2(\alpha F^2 + \beta F + \gamma), \qquad \frac{\partial \Lambda}{\partial F} = -\Lambda^2(2F + a + M_s(2\alpha F + \beta)),$$

and therefore one computes

$$\begin{aligned}
\frac{\partial f}{\partial M_s} &= -\mu F^2\Lambda^2(\alpha F^2+\beta F+\gamma),\\
\frac{\partial f}{\partial F} &= 2\mu F\Lambda-\mu F^2\Lambda^2(2F+a+M_s(2\alpha F+\beta))-\delta_F,\\
\frac{\partial^2 f}{\partial M_s^2} &= 2\mu F^2\Lambda^3(\alpha F^2+\beta F+\gamma)^2\\
\frac{\partial^2 f}{\partial M_s\partial F} &= -\mu F\Lambda^2(2(\alpha F^2+\beta F+\gamma)+F(2\alpha F+\beta)),\\
&\quad +2\mu F^2\Lambda^3(2F+a+M_s(2\alpha F+\beta))(\alpha F^2+\beta F+\gamma).
\end{aligned}$$

After straightforward but tedious computations, we find

$$\begin{aligned}
&\frac{\partial^2 f}{\partial M_s\partial F}f \;-\; \frac{\partial f}{\partial M_s}\frac{\partial f}{\partial F}\\
&= -\mu F^2\Lambda^3(2\mu F(\alpha F^2+\beta F+\gamma)+\mu F^2(2\alpha F+\beta))\\
&\quad +2\mu^2F^4\Lambda^4(2F+a+M_s(2\alpha F+\beta))(\alpha F^2+\beta F+\gamma)\\
&\quad +\delta_F F\Lambda^2\big(2\mu F(\alpha F^2+\beta F+\gamma)+\mu F^2(2\alpha F+\beta)\big)\\
&\quad -2\mu\delta_F F^3\Lambda^3(2F+a+M_s(2\alpha F+\beta))(\alpha F^2+\beta F+\gamma)\\
&\quad +\mu F^2\Lambda^2(\alpha F^2+\beta F+\gamma)(2\mu F\Lambda-\delta_F-\mu(2F+a+M_s(2\alpha F+\beta))\Lambda^2F^2)\\
&= \;\mu^2F^4\Lambda^4\Big((2F+a+M_s(2\alpha F+\beta))(\alpha F^2+\beta F+\gamma)-\frac{2\alpha F+\beta}{\Lambda}\Big)\\
&\quad +\mu F^2\Lambda^3\delta_F\times\Big(\frac{3\alpha F^2+2\beta F+\gamma}{\Lambda}\\
&\quad -2F(2F+a+M_s(2\alpha F+\beta))(\alpha F^2+\beta F+\gamma)\Big).
\end{aligned}$$

Using the expression of $1/\Lambda$ from (13), we get

$$\begin{aligned}
\frac{\partial^2 f}{\partial M_s\partial F}f-\frac{\partial f}{\partial M_s}\frac{\partial f}{\partial F} =&\mu^2F^4\Lambda^4((\beta-\alpha a)F^2+2\gamma F+a\gamma)\\
&+\mu F^2\Lambda^3\delta_F(-\alpha F^4+(a\alpha-2\beta)F^3-3\gamma F^2-a\gamma F))\\
&+\mu F^2\Lambda^3\delta_F M_s(\alpha F^2+\beta F+\gamma)(\gamma-\alpha F^2).
\end{aligned}$$

Finally, we obtain

$$\begin{aligned}
\mathfrak{U} &= \mu^2F^4\Lambda^4((\beta-\alpha a)F^2+2\gamma F+a\gamma)\\
&\quad +\mu F^3\Lambda^3\delta_F(-\alpha F^3+(a\alpha-2\beta)F^2-3\gamma F-a\gamma))\\
&\quad +\mu F^2\Lambda^3\delta_F M_s(\alpha F^2+\beta F+\gamma)(\gamma-\alpha F^2)\\
&\quad -\delta_s M_s 2\mu F^2\Lambda^3(\alpha F^2+\beta F+\gamma)^2.
\end{aligned}$$

Let us use that $2\delta_s>\delta_F$. Noting that $f\leqslant 0$ as a consequence of Lemma 4.10, which rewrites $\mu F\Lambda\leqslant\delta_F$, we obtain

$$\begin{aligned}
\mathfrak{U} &\leqslant -\mu^2F^4\Lambda^4\alpha a F^2+\mu F^3\Lambda^3\delta_F(-\alpha F^3+(a\alpha-\beta)F^2-\gamma F))\\
&\quad -\mu F^2\Lambda^3\delta_F M_s(\alpha F^2+\beta F+\gamma)(2\alpha F^2+\beta F).
\end{aligned}\tag{14}$$

We observe that this quantity is negative whenever $\beta > a\alpha$, which is the case since

$$a = \frac{K(\nu_E + \delta_E)}{\beta_E}, \quad \alpha = \frac{\delta_M \gamma_s}{K(1-\nu)\nu_E} \quad \text{and} \quad \beta = \frac{2\delta_M \gamma_s(\nu_E + \tau_E)}{(1-\nu)\beta_E \nu_E}.$$

(ii) Let I be an open interval of $\{R < -1/\lambda\}$. Let $\tau \in I$ be a local extremum of R. Therefore, we have $R'(\tau) = 0$. According to Lemma 4.11, one has $u^* = \overline{U}$ on I. Hence, from (12), one gets

$$-R'' = \left(\frac{\partial^2 f}{\partial M_s \partial F} f + \frac{\partial^2 f}{\partial M_s^2}(\overline{U} - \delta_s M_s) - \frac{\partial f}{\partial M_s}\frac{\partial f}{\partial F} \right) Q - \delta_s R'. \tag{15}$$

Recall that $\varepsilon \leqslant F(\cdot) \leqslant F(0) = \overline{F}$ according to Lemma 4.10. By using also Lemma 4.13, we get the existence of $C > 0$ independent of T and $\overline{U}$ such that

$$\begin{aligned} \frac{\partial^2 f}{\partial M_s^2}(F, M_s) &= \frac{2\nu(1-\nu)\beta_E^2 \nu_E^2 F^2 \delta_M^2 \gamma_s^2 (\frac{\beta_E F}{K} + \nu_E + \delta_E)}{\left((1-\nu)\nu_E \beta_E F + \delta_M \gamma_s M_s (\frac{\beta_E F}{K} + \nu_E + \delta_E)\right)^3} \\ &\geqslant \frac{2\nu(1-\nu)\beta_E^2 \nu_E^2 F^2 \delta_M^2 \gamma_s^2 (\nu_E + \delta_E)}{\left((1-\nu)\nu_E \beta_E \overline{F} + \delta_M \gamma_s C_0 (\frac{\beta_E \overline{F}}{K} + \nu_E + \delta_E)\right)^3} > C. \end{aligned} \tag{16}$$

A similar reasoning shows that all the other terms in (15) are uniformly bounded with respect to T and $\overline{U}$. Thus, we can find $\overline{U}$ (independent on T) large enough such that the right-hand side is positive. Then, $R''(\tau) < 0$, which implies R admits a local maximum at τ.

□

Lemma 4.16. *Let us make the same assumption as in Proposition 2.3. Consider the dynamical system ($\mathcal{S}_1$) with $u = 0$, i.e.*

$$\frac{dF}{dt} = f(F, M_s), \qquad \frac{dM_s}{dt} = -\delta_s M_s,$$

where f is given in (1). If $F'(\tau) \geqslant 0$ then, for all $t > \tau$, we have $F'(t) \geqslant 0$.

Proof. This is a consequence of the fact that the set

$$\mathcal{E} := \{(F, M_s) \in \mathbb{R}_+ \times (0, +\infty), \text{ such that } f(F, M_s) \geqslant 0\}$$

is stable by the aforementioned dynamical system. Indeed, on $\mathbb{R}_+^2$, $f(F, M_s) \geqslant 0$ if, and only if, $0 \leqslant M_s \leqslant \phi(F)$, where the function ϕ is obtained by solving the implicit equation $f(F, \phi(F)) = 0$. If at some time $\tau > 0$, a trajectory crosses the part of the boundary of $\mathcal{E}$ defined by the implicit equation, the vector field defining the right-hand side of the differential system reads

$$[f(F(\tau), M_s(\tau)), -\delta_s M_s(\tau)]^\top = [0, -\delta_s M_s(\tau)]^\top.$$

This field is vertically directed, and it points inward for $\mathcal{E}$. The conclusion is similar on the other parts of the boundary of $\mathcal{E}$, whence the stability of this zone. □

Proof of Theorem 3.3. We separately deal with the characterization and uniqueness properties minimizers.

Step 1: characterization of optimal controls. Notice first that the sets $\{R > -1/\lambda\}$ and $\{R < -1/\lambda\}$ are open, as inverse images of open sets by continuous functions (and thus contain an interval). By combining Lemma 4.15, Lemma 4.12, and the fact that R is continuous with $R(T) = 0$, we get the existence of $(t_0, t_1) \in [0, T]^2$ such that $0 \leq t_0 \leq t_1 < T$ and

- $R > -1/\lambda$ on $(0, t_0)$ or $R < -1/\lambda$ on $(0, t_0)$;
- $R = -1/\lambda$ on (t_0, t_1);
- $R > -1/\lambda$ on (t_1, T).

Combined with Lemmas 4.11 and 4.14, we deduce the form of the optimal control in Theorem 3.3.

Let us now prove that for T large enough, we have $t_0 > 0$ and $u^* = 0$ on $(0, t_0)$. Assume by contradiction that $u^* = \overline{U}$ on $(0, t_0)$ or $t_0 = 0$. Recall that, according to Lemma 4.8, we have

$$J^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon} = \int_0^T u^*(t)\, dt \leqslant \overline{U}\, \overline{T}.$$

However, combining the expression of u^* given in (7) with the estimates (14) and (16) show that

$$u^*(t) \geqslant \tilde{C} > 0, \qquad \text{on } (t_0, t_1),$$

where $\tilde{C}$ does not depend on neither T nor $\overline{U}$. Without loss of generality, we can assume that $\tilde{C} < U^*$. Hence, one has

$$\int_0^{t_1} u^*(t)\, dt \geqslant t_1 \tilde{C}$$

(since $u^* = \overline{U} > U^* > \tilde{C}$ on $(0, t_0)$). It follows that $t_1 \leqslant \overline{UT}/\tilde{C}$ is uniformly bounded with respect to T. On (t_1, T), one has $u^* = 0$. Let $F_{\overline{U}}$ denote the trajectory associated to the control choice $u_{\overline{U}} = \overline{U}\mathbb{1}_{[0,\overline{UT}/\tilde{C}]}$. According to the monotonicity property stated in Lemma 4.7, one has $F \geqslant F_{\overline{U}}$ in $\mathbb{R}_+$ since $u \leqslant u_{\overline{U}}$. Furthermore, according to Proposition 2.3, $F_{\overline{U}}(T)$ converges to the steady state $\overline{F}$ as $T \to +\infty$. Let $T_{\overline{U}} > 0$ be given such that $F_{\overline{U}}(t) \geqslant (\varepsilon + \overline{F})/2 > \varepsilon$ in $(T_{\overline{U}}, +\infty)$. If $T > T_{\overline{U}}$, one thus has $F(T) > \varepsilon$ which contradicts the fact that $F(T) = \varepsilon$ (see Lemma 4.9).

Hence, there exists $T^* > 0$ large enough such that, for $T > T^*$, we have $t_0 > 0$, and $u^* = 0$ on $(0, t_0)$. Necessarily, $t_1 > t_0$, otherwise $u^* = 0$ a.e. on $(0, T)$ which is not possible, since in this case $F(t) = \overline{F} > \varepsilon$ on $(0, T)$.

Let us notice that if $T > T^*$, we have

$$F'(T) \geqslant 0. \tag{17}$$

Indeed, since $T > T^*$, we have seen that the optimal control has the form $u^* = u^*\mathbb{1}_{(t_0,t_1)}$ for $0 < t_0 < t_1 < T$. If $F'(T) < 0$, then there exists $T_1 > T$ such that $F(T_1) < \varepsilon$ and $T_1 - T < t_0$. Then, by taking $u = u^*\mathbb{1}_{(t_0-T_1+T,t_1-T_1+T)}$, we obtain a control on $(0, T)$ such that $J(u) = J(u^*)$ and $F(T) < \varepsilon$. However,this is not possible (see Lemma 4.9). Thus, $F'(T) \geqslant 0$.

Finally, let us show the claimed stationarity property of optimal values. To this aim, let us consider $T_1 > T_2 > T^*$. Since $T \mapsto J_{T,\overline{U},\varepsilon}$ is non-increasing, then $J^{(\mathcal{S}_1)}_{T_1,\overline{U},\varepsilon} \leqslant J^{(\mathcal{S}_1)}_{T_2,\overline{U},\varepsilon}$. Let us show that $J^{(\mathcal{S}_1)}_{T_1,\overline{U},\varepsilon} = J^{(\mathcal{S}_1)}_{T_2,\overline{U},\varepsilon}$. By contradiction, assume that $J^{(\mathcal{S}_1)}_{T_1,\overline{U},\varepsilon} < J^{(\mathcal{S}_1)}_{T_2,\overline{U},\varepsilon}$. Let us denote by u_1^* (resp. u_2^*), the optimal solution on $(0, T_1)$ (resp. $(0, T_2)$). Then, from the results above, there exist $t_0^{(1)} < t_1^{(1)}$ and $t_0^{(2)} < t_1^{(2)}$ such that $u_1^* = u^*(\cdot - t_0^{(1)})\mathbb{1}_{(t_0^{(1)},t_1^{(1)})}$ and $u_2^* = u^*(\cdot - t_0^{(2)})\mathbb{1}_{(t_0^{(2)},t_1^{(2)})}$ with the same expression of u^* given in (7). From

$$\int_{t_0^{(1)}}^{t_1^{(1)}} u^*(t - t_0^{(1)})\, dt = J^{(\mathcal{S}_1)}_{T_1,\overline{U},\varepsilon} < J^{(\mathcal{S}_1)}_{T_2,\overline{U},\varepsilon} = \int_{t_0^{(2)}}^{t_1^{(2)}} u^*(t - t_0^{(2)})\, dt,$$

we deduce that $t_1^{(1)} - t_0^{(1)} < t_1^{(2)} - t_0^{(2)}$. Notice that, denoting F_1, resp. F_2, the solution to $(\mathcal{S}_1)$ with $u = u_1^*$, resp. $u = u_2^*$, we have $F_1(t + t_0^{(1)}) = F_2(t + t_0^{(2)})$ for each $t \in (0, t_1^{(1)} - t_0^{(1)})$. Moreover, from

the above remark, we have that $F_1'(T_1) \geqslant 0$, $F_2'(T_2) \geqslant 0$, and (using Lemma 4.16) that $F_2'(t) > 0$ for $t > T_2$. Then, $F_1(t_1^{(1)}) = F_2(t_1^{(1)} + t_0^{(2)} - t_0^{(1)}) > F_2(t_1^{(2)})$. Since $t \mapsto F_1(t + t_1^{(1)})$ and $t \mapsto F_2(t + t_1^{(2)})$ verify the same dynamical system on $(0, T_1 - t_1^{(1)})$, we deduce that $F_1(t + t_1^{(1)}) > F_2(t + t_1^{(2)})$ which implies in particular that $F_1(T_1) > F_2(T_1 + t_1^{(2)} - t_1^{(1)}) \geqslant \varepsilon$. This is in contradiction with the fact that $F_1(T_1) = \varepsilon$ (see Lemma 4.9).

It remains to show that F has its minimal value at T and that $F'(T) = 0$. Let us extend the definition of F to $\mathbb{R}_+$ by setting $u^* = 0$ in $(T, +\infty)$. We already know that F is non-increasing on $[0, T]$ according to Lemma 4.10 and therefore, $F'(T) \leq 0$, which leads to the conclusion since $F'(T) \geqslant 0$ (see (17)).

Step 2: uniqueness property. To conclude the proof, let us prove the uniqueness of the optimal control if $T > T^*$. We will use a constructive argument which is also used to derive an algorithm for solving numerically this problem in Section 3.2.

For $(t_0, t_1) \in [0, T]^2$ such that $t_0 \leqslant t_1$, introduce $(F_{(t_0,t_1)}, M_{s(t_0,t_1)})$ as the solution of the Cauchy problem

$$\begin{cases} \dfrac{d}{dt}\begin{pmatrix} F \\ M_s \end{pmatrix} = \begin{pmatrix} f(F, M_s) \\ u - \delta_s M_s \end{pmatrix} \quad \text{in } [0, +\infty) \\ \text{with } u_{(t_0,t_1)} = \dfrac{\frac{\partial f}{\partial M_s}(F,M_s)\frac{\partial f}{\partial F}(F,M_s) + \frac{\partial^2 f}{\partial M_s^2}(F,M_s)\delta_s M_s - \frac{\partial^2 f}{\partial M_s \partial F}(F,M_s) f(F,M_s)}{\frac{\partial^2 f}{\partial M_s^2}(F,M_s)} \mathbb{1}_{[t_0,t_1]}, \end{cases}$$

complemented with the initial conditions $F(0) = \overline{F}$ and $M_s(0) = 0$.

Let us first highlight that, with the notations of Theorem 3.3, it is enough to consider the case where t_0 is equal to 0.

Lemma 4.17. *Under the assumptions of Theorem 3.3 and if $T > T^*$, $u_{(t_0,t_1)}$ solves Problem $(\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$ if and only if the function $\tilde{u}$ given by $\tilde{u}(t) = u_{(t_0,t_1)}\big|_{(t_0,T)}(t - t_0)$ solves Problem $(\mathcal{P}^{(\mathcal{S}_1)}_{T-t_0,\overline{U},\varepsilon})$.*

Let us prove this lemma. The characterization of optimal controls in the previous step shows that any solution of Problem $(\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$ is of the form $u_{(t_0,t_1)}$ for some $0 \leq t_0 < t_1 \leq T$. Observe that

$$\int_0^T u_{(t_0,t_1)}(t)\, dt = \int_0^{T-t_0} \tilde{u}(t)\, dt$$

and furthermore, denoting respectively by $F_{u_{(t_0,t_1)}}$ and $F_{\tilde{u}}$ the trajectories associated to $u_{(t_0,t_1)}$ and $\tilde{u}$, one has by construction $F_{u_{(t_0,t_1)}}(T) = F_{\tilde{u}}(T - t_0) = \varepsilon$. Since $T \in (T^*, +\infty) \mapsto J^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon}$ is nondecreasing, it is constant on $(T - t_0, T)$ and it follows that $\tilde{u}$ necessarily solves Problem $(\mathcal{P}^{(\mathcal{S}_1)}_{T-t_0,\overline{U},\varepsilon})$. The proof of converse sense is exactly similar (and left to the reader). This ends the proof of Lemma 4.17.

Let us argue by contradiction to prove the uniqueness of optimizers, assuming that Problem $(\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$ has two solutions $u_{(t_0,t_1)}$ and $u_{(t_0',t_1')}$. According to Lemma 4.17 above, $u_{(0,t_1-t_0)}$ and $u_{(0,t_1'-t_0')}$ solve Problems $(\mathcal{P}^{(\mathcal{S}_1)}_{T-t_0,\overline{U},\varepsilon})$ and $(\mathcal{P}^{(\mathcal{S}_1)}_{T-t_0',\overline{U},\varepsilon})$ respectively. Without loss of generality, assume that $t_1' - t_0' \leqslant t_1 - t_0$. Since these controls are non-negative and $\int_{(0,T)} u_{(0,t_1-t_0)} = \int_{(0,T)} u_{(0,t_1'-t_0')}$, we infer that the function

$$\frac{\frac{\partial f}{\partial M_s}(F,M_s)\frac{\partial f}{\partial F}(F,M_s) + \frac{\partial^2 f}{\partial M_s^2}(F,M_s)\delta_s M_s - \frac{\partial^2 f}{\partial M_s \partial F}(F,M_s) f(F,M_s)}{\frac{\partial^2 f}{\partial M_s^2}(F,M_s)}$$

vanishes on $(t_1 - t_0, t_1' - t_0')$. If $t_1' - t_0' < t_1 - t_0$, a standard regularity argument for Cauchy problems yields that $u_{(0,t_1-t_0)}$ is real analytic on $(0, t_1 - t_0)$, and vanishes hence identically on $(0, T)$. Therefore, $F_{u_{(0,t_1-t_0)}}(\cdot) = \overline{F}$ on $\mathbb{R}_+$, which is impossible since $F_{u_{(0,t_1-t_0)}}(T - t_0) = \varepsilon$. Thus $t_1' - t_0' = t_1 - t_0$ and $F_{u_{(0,t_1-t_0)}} = F_{u_{(0,t_1'-t_0')}}$ on $\mathbb{R}_+$. The function $F_{u_{(0,t_1-t_0)}}$ is C^1 on $\mathbb{R}_+$ and one has $\lim_{t\to+\infty} F_{u_{(0,t_1-t_0)}}(t) = \overline{F}$, according to Proposition 2.3 and since the persistence equilibrium is the only one not being unstable. Let us denote by t_2 the time at which $F_{u_{(0,t_1-t_0)}}$ admits its first local minimum. Thanks to Lemma 4.16, $F_{u_{(0,t_1-t_0)}}$ decreases on $(0, t_2)$ and increases on $(t_2, +\infty)$. Since $u_{(0,t_1-t_0)}$ solves Problems $(\mathcal{P}^{(\mathcal{S}_1)}_{T-t_0,\overline{U},\varepsilon})$ and $(\mathcal{P}^{(\mathcal{S}_1)}_{T-t_0',\overline{U},\varepsilon})$, one has $F'_{u_{(0,t_1-t_0)}}(T-t_0) = F'_{u_{(0,t_1-t_0)}}(T-t_0') = F'(t_2) = 0$. We deduce that $T - t_0 = T - t_0' = t_2$, which shows the uniqueness of t_0 and t_1. □

4.3 Proof of Property 3.5

We recall that, for each $\tau_1[0, T)$, u_{τ_1} and F_{τ_1}, denote the solution to System (8), in particular u_∞ and F_∞ are the solutions to (8) for $\tau_1 = \infty$. Notice first that F^{τ_1} is C^1 on $(0, +\infty)$ for each $\tau_1 \in [0, T)$. Denote by $t_{\max} \in (0, +\infty]$ the minimal time t at which it holds either $(F^\infty)'(t) > 0$ or $u_\infty(t) < 0$ (we do not know a priori if u_∞ can be negative or not since u_∞ is the solution to System (8), not to the optimal control problem). Its existence results from the form of optimal controls (see Theorem 3.3). Let us first prove that $t_{\max} = +\infty$. Assume by contradiction that $t_{\max} < +\infty$. According to Proposition 2.3 and since the persistence equilibrium is the only one being not unstable, one has $\lim_{t\to+\infty} F^{t_{\max}}(t) = \overline{F}$. Let τ_2 be the time at which $F^{t_{\max}}$ admits its first local minimum. According to Lemma 4.16, $F^{t_{\max}}$ is decreasing on $(0, \tau_2)$ and increasing on (τ_2, ∞). Hence, there exists $\delta > 0$ such that $F^{t_{\max}} \geqslant \delta$. To reach a contradiction, let us consider a particular choice of ε, such that $\varepsilon \in (0, \delta)$. Let $T > 0$ be large enough so that Problem $(\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$ is well-posed (see Theorem 3.3) and let $u_{(t_0,t_1)}$ be its unique solution. According to Lemma 4.17, $u_{t_1-t_0}$ solves Problem $(\mathcal{P}^{(\mathcal{S}_1)}_{T-t_0,\overline{U},\varepsilon})$. Since $u_{t_1-t_0}$ is positive and $(F^{t_1-t_0})' > 0$ on $(0, t_1 - t_0)$, one has $t_1 - t_0 \leqslant t_{\max}$. By Lemma 4.7, one has $F^{t_1-t_0} \geqslant F^{t_{\max}} > \delta$ on $\mathbb{R}_+$, which is in contradiction with the fact $F^{t_1-t_0}(T) = \varepsilon$. Thus $t_{\max} = \infty$.

Item (i) can be obtained with the same reasoning as above and the fact that $t_{\max} = +\infty$. Property (ii) is a consequence of Lemma 4.7 and again of the equality $t_{\max} = +\infty$. Finally, the proof of the last claim (iii), establishing connections between Problem $(\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$ under its general form and the control u_{τ_1}, follows from Lemma 4.17, used to get the uniqueness of optimal controls in the proof of Theorem 3.3.

5 Comments on the optimal control problem and modeling issues

In this section, we introduce two other possible choices of functionals to minimize and compare it to the one analyzed in the previous sections.

5.1 L^2 functional

Consider the optimal control problem

$$\inf_{u\in\mathcal{U}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon}} \tilde{J}(u), \qquad (\tilde{\mathcal{P}}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$$

where the functional $\tilde{J}$ stands for the square of the L^2-norm of released mosquitoes over the horizon of time T, namely

$$\tilde{J}(u) := \int_0^T u(t)^2 dt$$

and $\mathcal{U}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon}$ is defined by (5).

Theorem 5.1. *Let $\varepsilon \in (0, \overline{F})$. For any $\overline{U} > U^*$ (defined by (3)), there exists a minimal time $\overline{T}(\overline{U}) > 0$ such that for all $T \geqslant \overline{T}(\overline{U})$, the set $\mathcal{U}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon}$ is nonempty and the optimal control problem $(\tilde{\mathcal{P}}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$ has a solution u^*. Moreover, for $\overline{U} > U^*$ and $T > \overline{T}(\overline{U})$ and large enough, there exists $\mu > 0$ and $t_0 \in [0, T)$ such that*

$$u^* = \begin{cases} \overline{U} & \text{on } [0, t_0) \\ -\mu R, & \text{on } [t_0, T] \end{cases}$$

where R solves the dual system (11). Moreover $u^*(T) = 0$ and the mappings $T \in [\overline{T}, +\infty) \mapsto \tilde{J}^{(\mathcal{S}_1)}_{T,\overline{U},C}$ and $\overline{U} \in (U^*, +\infty) \mapsto \tilde{J}^{(\mathcal{S}_1)}_{T,\overline{U},C}$ are non-increasing.

Remark 5.2. *Similarly to the L^1 case, we have reduced the infinite dimensional control problem to finite dimensional one, and we therefore only need to determine μ and the value of Q and R at time 0 to see (F, M_s, Q, R) as the solution of a well-posed Cauchy problem.*

To prove Theorem 5.1, we first need the equivalent of Lemma 4.11.

Lemma 5.3. *Consider u^* a solution to Problem $(\tilde{\mathcal{P}}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$. Then there exists $\lambda > 0$ such that*

$$\begin{cases} \text{a.e. on } \{u^* = 0\}, \text{ one has } u^*(t) + \lambda R(t) \geqslant 0, \\ \text{a.e. on } \{0 < u^* < \overline{U}\}, \text{ one has } u^*(t) + \lambda R(t) = 0, \\ \text{a.e. on } \{u^* = \overline{U}\}, \text{ one has } u^*(t) + \lambda R(t) \leqslant 0. \end{cases}$$

The proof is similar to Lemma 4.11 and will be omitted.

Proof of Theorem 5.1. From (11), we deduce that

$$R' = -\frac{\partial f}{\partial M_s} Q + \delta_s R > \delta_s R,$$

where we use the fact that $Q > 0$ and $\partial f / \partial M_s < 0$. Since one has $R(T) = 0$, we infer that $R < 0$ in $[0, T)$ by using a standard Gronwall argument (see e.g. [20]). According to Lemma 5.3, one has $\{u^* = 0\} = \emptyset$.

According to Lemma 4.13, since $0 \leqslant \partial f / \partial M_s(M_s, F) \leqslant \mathrm{O}(|F|)$, one infers that R is uniformly bounded on $[0, T]$, by a constant that does not depend on $\overline{U}$ and T. Then, by adapting the proof of Lemma 4.15, one shows that for $\overline{U}$ large enough, on any open interval of the open set $\{\overline{U} + \lambda R < 0\}$, any local extremum of R is a maximum. It follows that $\{\overline{U} + \lambda R < 0\}$ has at most one connected component of the form $(0, t_0)$, which leads to the conclusion. □

We provide on Figure 7 the solutions of Problem $(\tilde{\mathcal{P}}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$ with $T = 200$, $\overline{U} = 4000$, $\nu_E = 0.05$ and $\varepsilon = \overline{F}/4$. We recover the theoretical result above, namely that the optimal control u^* is positive, continuous and $u^*(T) = 0$.

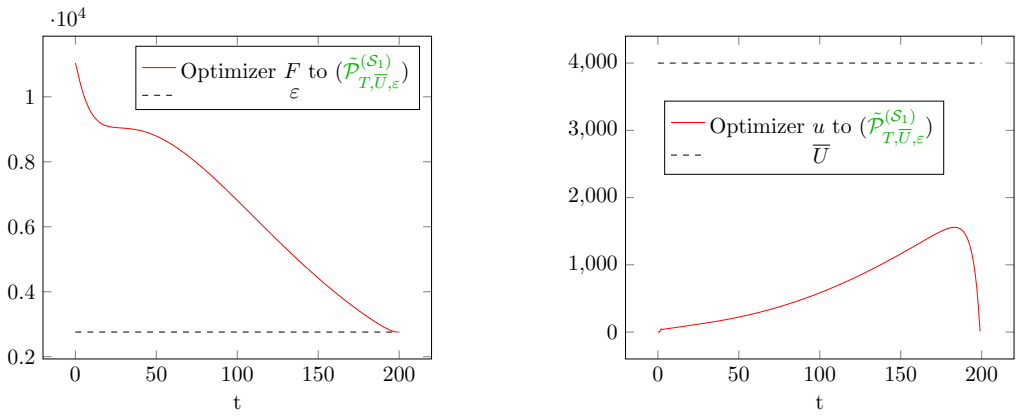

Figure 7: Solution of the optimal control problems $(\tilde{\mathcal{P}}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$ with $T = 200$, $\overline{U} = 4000$, $\nu_E = 0.05$ and $\varepsilon = \overline{F}/4$.

Remark 5.4. *In optimal control theory, the L^2-norm is often preferred to the L^1-norm for differentiability issues. However, from a biological point of view, the L^1-norm is more relevant since it stands for the amount of individuals. Moreover, as it can be seen on Figure 3, the optimal control for the L^1-norm is sparse unlike the one for the L^2-norm, which is interesting from a practical point of view.*

5.2 Dual optimal control problem

Consider the optimal control problem

$$\inf_{u \in \mathcal{U}^{(\mathcal{S}_1)}_{T,\overline{U},C}} \hat{J}(u), \qquad (\hat{\mathcal{P}}^{(\mathcal{S}_1)}_{T,\overline{U},C})$$

where the functional $\hat{J}$ stands for the total number of eggs and females (with some weights) at time T, namely

$$\hat{J}(u) := F(T),$$

where F solves System $(\mathcal{S}_1)$ associated to the control u and $\mathcal{U}^{(\mathcal{S}_1)}_{T,\overline{U},C}$ is the set of admissible controls, chosen so that:

- the rate of sterile male mosquito release is non-negative, uniformly bounded by a positive constant $\overline{U}$;
- the total number of released sterilized males over the time interval $(0, T)$ is assumed to be lower than C.

Hence, $\mathcal{U}^{(\mathcal{S}_1)}_{T,\overline{U},C}$ is defined by

$$\mathcal{U}^{(\mathcal{S}_1)}_{T,\overline{U},C} := \left\{ u \in L^\infty(0,T) : 0 \leqslant u \leqslant \overline{U} \text{ a.e. in } (0,T), \int_0^T u(t)\,dt \leqslant C \right\}.$$

In [1], a similar optimal control problem has been considered. Problem $(\hat{\mathcal{P}}^{(\mathcal{S}_1)}_{T,\overline{U},C})$ can be seen as a dual version of $(\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$ in the following sense: let $u^* \in \mathcal{U}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon}$ be a solution to Problem $(\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$

for some given T, $\overline{U}$ and ε. Then, the control u^* is a solution to Problem $(\hat{\mathcal{P}}^{(\mathcal{S}_1)}_{T,\overline{U},C})$ for the parameter choice $C := \int_0^T u^*(t)\,dt$. Indeed, assume by contradiction that there exists $u \in \mathcal{U}^{(\mathcal{S}_1)}_{T,\overline{U},C}$ such that

$$\hat{J}(u) < \hat{J}(u^*),$$

that is $F(T) < F^*(T) = \varepsilon$, where (F, M_s) and (F^*, M_s^*) are the solutions to system $(\mathcal{S}_1)$ associated to u and u^* respectively. By mimicking the argument provided in the proof of Lemma 4.3, we reach a contradiction, which shows that u^* is a solution to Problem $(\hat{\mathcal{P}}^{(\mathcal{S}_1)}_{T,\overline{U},C})$.

Respectively, let $\hat{u}^* \in \mathcal{U}^{(\mathcal{S}_1)}_{T,\overline{U},C}$ be an optimizer to Problem $(\hat{\mathcal{P}}^{(\mathcal{S}_1)}_{T,\overline{U},C})$ for some given T, $\overline{U}$ and C. By using the same argument, one shows that $\hat{u}^*$ is an optimizer to Problem $(\mathcal{P}^{(\mathcal{S}_1)}_{T,\overline{U},\varepsilon})$ for the parameter choice $\varepsilon := F^*(T)$.

6 Conclusion

In this paper, we have determined the optimal release function which minimizes the number of sterilized males needed when performing the sterile insect technique (SIT) to reduce the size of a population of mosquitoes to a given value. Starting from a differential system modeling the dynamics of the mosquito population, we simplify it to obtain a reduced system, which is a good approximation, and for which we are able to compute precisely the optimal solution. These theoretical results are illustrated thanks to some numerical simulations. Notice that once the form of the theoretical solution is known, efficient algorithms may be designed to compute quickly the numerical solution.

Obviously, when the final number of mosquitoes is fixed, there is a minimal time to perform the releases in order to reach this value. Interestingly, the number of sterilized males needed is non-increasing with respect to the time of the experiment, meaning that the longer the duration of the experiment, the lower the number of sterilized males. However, there is a maximal time above which the minimal number of sterilized males needed is stationary. In this case, the optimal release function is given by a singular arc sandwiched between two regions where it is zero. The knowledge of the existence of this time may be interesting for practical applications since for larger time the number of sterilized males stays constant.

Thanks to our results, we are able to give a precise description of the temporal distribution of the releases to optimize given scenarios. A natural extension of this work is to add the spatial distribution of mosquitoes since it may have a big impact on the success of the SIT (we refer the interested reader to [2, 16] for some simple spatial models). In particular, most experiments have been performed in isolated regions to avoid re-invasion from the outside. But even in isolated regions the question of knowing where to perform the releases to have the best efficiency of the SIT is still open.

Appendix

A Mathematical properties of the dynamical systems

Proof of Proposition 2.1

Let us assume that $u(\cdot) = 0$. Then, the equilibria $(\overline{E}, \overline{M}, \overline{F}, \overline{M}_s)$ of System $(\mathcal{S}_2)$ solve

$$0 = \beta_E \overline{F}\left(1 - \frac{\overline{E}}{K}\right) - \big(\nu_E + \delta_E\big)\overline{E} = (1-\nu)\nu_E \overline{E} - \delta_M \overline{M}$$
$$= \nu\nu_E \overline{E}\frac{\overline{M}}{\overline{M} + \gamma_s \overline{M}_s} - \delta_F \overline{F} = -\delta_s \overline{M}_s.$$

We thus infer that $\overline{M}_s = 0$ and

$$0 = \beta_E \overline{F}\left(1 - \frac{\overline{E}}{K}\right) - \big(\nu_E + \delta_E\big)\overline{E} = (1-\nu)\nu_E \overline{E} - \delta_M \overline{M} = \nu\nu_E \overline{E} - \delta_F \overline{F}.$$

Then, $(0,0,0,0)$ is an equilibrium, and the only non-zero equilibrium is

$$\overline{E} = K\left(1 - \frac{\delta_F\big(\nu_E + \delta_E\big)}{\beta_E \nu\nu_E}\right), \quad \overline{M} = \frac{(1-\nu)\nu_E}{\delta_M}\overline{E}, \quad \overline{F} = \frac{\nu\nu_E}{\delta_F}\overline{E}, \quad \overline{M}_s = 0$$

whence (2).

Let us show that $(0,0,0,0)$ is unstable. Using that $M_s(t) = e^{-\delta_s t} M_s(0)$ for $t \geqslant 0$, we deduce that $M(t) \geqslant e^{-\delta_M t} M(0)$ for $t \geqslant 0$ according to $(\mathcal{H})$, and it follows that for any $\epsilon > 0$, there exists $t^* > 0$ such that $\gamma_s M_s(t) < \epsilon M(t)$ for all $t \geqslant t^*$. By using a standard comparison principle, we get that for all $t \geqslant t^*$

$$(E(t), F(t)) \geqslant (E_1(t), F_1(t)),$$

the inequality being understood component by component, where, (E_1, F_1) solves

$$\begin{cases} \dfrac{dE_1}{dt} = \beta_E F_1\left(1 - \dfrac{E_1}{K}\right) - \big(\nu_E + \delta_E\big)E_1, \quad t \geqslant t^* \\ \dfrac{dF_1}{dt} = \dfrac{\nu\nu_E}{1+\epsilon}E_1 - \delta_F F_1 \end{cases} \tag{18}$$

complemented with the initial data $(E_1(t^*), F_1(t^*)) = (E(t^*), F(t^*))$. An easy computation yields that the Jacobian matrix of System (18) at $(0,0)$ is

$$\begin{pmatrix} -\nu_E - \delta_E & \beta_E \\ \frac{\nu\nu_E}{1+\epsilon} & -\delta_F \end{pmatrix},$$

whose determinant expands as $\delta_F(\nu_E + \delta_E) - \nu\beta_E\nu_E + \mathrm{O}(\epsilon)$. According to $(\mathcal{H})$, we infer that the Jacobian matrix has a positive root whenever ϵ is chosen small enough, which leads to the conclusion.

Let us now investigate the stability of $(\overline{E}, \overline{M}, \overline{F}, 0)$ for System $(\mathcal{S}_2)$. Easy computations yield

that the Jacobian matrix of System ($\mathcal{S}_2$) at $(\overline{E},\overline{M},\overline{F},0)$ reads

$$\begin{pmatrix} -\frac{\beta_E}{K}\overline{F}-(\nu_E+\delta_E) & 0 & \beta_E\left(1-\frac{\overline{E}}{K}\right) & 0 \\ (1-\nu)\nu_E & -\delta_M & 0 & 0 \\ \nu\nu_E & 0 & -\delta_F & -\frac{\gamma_s\nu\nu_E\overline{E}}{\overline{M}} \\ 0 & 0 & 0 & -\delta_s \end{pmatrix}$$

$$= \begin{pmatrix} -\frac{\nu\nu_E\beta_E}{\delta_F} & 0 & \frac{\delta_F(\nu_E+\delta_E)}{\nu\nu_E} & 0 \\ (1-\nu)\nu_E & -\delta_M & 0 & 0 \\ \nu\nu_E & 0 & -\delta_F & -\frac{\gamma_s\nu\delta_M}{1-\nu} \\ 0 & 0 & 0 & -\delta_s \end{pmatrix}.$$

so that the four eigenvalues are $-\delta_s$, $-\delta_M$, and the two (complex conjugate) roots of the polynomial

$$P = X^2 + \left(\frac{\nu\nu_E\beta_E}{\delta_F}+\delta_F\right)X + (\nu\nu_E\beta_E - \delta_F(\nu_E+\delta_E)),$$

which have a negative real part under ($\mathcal{H}$). It follows that $(\overline{E},\overline{M},\overline{F},0)$ is locally asymptotically stable.

Let us now prove the second part of the proposition. We first notice that the set $[0,K]\times\mathbb{R}_+^3$ is stable whenever u is non-negative. We claim that System ($\mathcal{S}_2$) is monotone on this set (see Lemma 4.1). If u belongs to $L^\infty(0,T,\mathbb{R}_+)$, then $(0,0,0,0)$ is obviously a subsolution of System ($\mathcal{S}_2$) whereas $(\overline{E},\overline{M},\overline{F},\|u\|_\infty/\delta_s)$ is a supersolution. A comparison argument allows us to conclude that $[0,\overline{E}]\times[0,\overline{M}]\times[0,\overline{F}]\times\mathbb{R}_+$ is stable. Furthermore, if all initial data are positive, then so are the functions E, M, F and M_s, and we get that

$$E(t)\geqslant e^{-(\nu_E+\delta_E)t}E(0),\quad M(t)\geqslant e^{-\delta_M t}M(0),\quad F(t)\geqslant e^{-\delta_F t}F(0)$$

for all $t\geq 0$, implying that these quantities cannot vanish.

Finally, let us consider the case where $u(\cdot)=\overline{U}$, where $\overline{U}>U^*$. In that case, the non-zero equilibria of System ($\mathcal{S}_2$) solve

$$\overline{M}_s=\frac{\overline{U}}{\delta_S},\quad \overline{M}=\frac{(1-\nu)\nu_E}{\delta_M}\overline{E},\quad \overline{F}=\frac{\nu\nu_E}{\delta_F}\frac{\overline{E}\,\overline{M}}{\overline{M}+\gamma_s\overline{M}_s},$$

and

$$0=\beta_E\overline{F}\left(1-\frac{\overline{E}}{K}\right)-(\nu_E+\delta_E)\overline{E}$$

plugging the three first above equalities into this latter equation, we get that $\overline{E}$ satisfies

$$\overline{E}=0$$

or

$$\frac{\beta_E\nu(1-\nu)\nu_E^2}{\delta_F\delta_M K}\overline{E}^2-\frac{\beta_E\nu(1-\nu)\nu_E^2}{\delta_F\delta_M}\left(1-\frac{\delta_F(\nu_E+\delta_E)}{\beta_E\nu\nu_E}\right)\overline{E}+\frac{\gamma_s(\nu_E+\delta_E)}{\delta_s}\overline{U}=0.$$

One easily checks that this second order polynomial with unknown $\overline{E}$ does not have any real solutions if $\overline{U}>U^*$. In this case, the only equilibrium is the extinction equilibrium. We infer that any non-negative solution to the Cauchy problem converges to this unique steady state. □

Proof of Proposition 2.3

Let us assume that $u(\cdot) = \overline{U}$. The equilibria are obtained by solving the system

$$0 = f(\overline{F}, \overline{M}_s) = \overline{U} - \delta_s \overline{M}_s,$$

which is equivalent to $\overline{M}_s = \overline{U}/\delta_s$ and

$$\overline{F}\left(\nu(1-\nu)\beta_E^2\nu_E^2\overline{F} - \delta_F\big(\frac{\beta_E\overline{F}}{K} + \nu_E + \delta_E\big)\Big((1-\nu)\nu_E\beta_E\overline{F} \right.$$
$$\left. +\delta_M\gamma_s\overline{M}_s\big(\frac{\beta_E\overline{F}}{K} + \nu_E + \delta_E\big)\Big)\right) = 0.$$

It follows that if $\overline{U} = 0$, then there are exactly two different solutions for this equation, namely $\overline{F} = 0$ or $\overline{F}$ given by (2). A straightforward computation yields that if $\overline{U} > U^*$, then the equation above has no positive solution and furthermore, $f(F, \overline{M}_s) < 0$ for every $F > 0$. Therefore, we infer that any non-negative solution converges to the steady state $(0, \overline{M}_s)$ if $\overline{U} > U^*$.

Let us assume that $u(\cdot) = 0$. Using that

$$M_s(t) = e^{-\delta_s t}M_s(0), \quad F(t) \geqslant e^{-\delta_F t}F(0)$$

for $t \geqslant 0$ and the fact that $\delta_s > \delta_F$, we get that, for all $\eta > 0$, there exists $t^* > 0$ such that

$$\delta_M\gamma_s\big(\frac{\beta_E F(t)}{K} + \nu_E + \delta_E\big)M_s(t) < \eta(1-\nu)\nu_E\beta_E F(t) \quad \text{for all } t \geqslant t^*.$$

It follows that

$$f(F(t), M_s(t)) > \tilde{f}(F(t)) := \frac{\nu\beta_E\nu_E F(t)}{\big(\frac{\beta_E F(t)}{K} + \nu_E + \delta_E\big)(1+\eta)} - \delta_F F(t),$$

whenever $t \geqslant t^*$. We conclude by observing that $\tilde{f}'(0) > 0$ for η small enough, so that we get the instability of the equilibrium $(0, 0)$.

Finally, let us investigate the stability of the persistence steady state $(\overline{F}, 0)$ of System $(\mathcal{S}_1)$. One computes

$$\begin{aligned}
&\tfrac{\partial f}{\partial F}(\overline{F}, 0) \\
&= \nu(1-\nu)\beta_E^2\nu_E^2\frac{2\overline{F}\Big(\frac{\beta_E\overline{F}}{K}+\nu_E+\delta_E\Big)(1-\nu)\nu_E\beta_E\overline{F} - \overline{F}^2(1-\nu)\nu_E\beta_E\Big(2\frac{\beta_E\overline{F}}{K}+\nu_E+\delta_E\Big)}{\Big(\frac{\beta_E\overline{F}}{K}+\nu_E+\delta_E\Big)^2\Big((1-\nu)\nu_E\beta_E\overline{F}\Big)^2} - \delta_F \\
&= \delta_F\Big(\tfrac{1}{\mathcal{R}_0} - 1\Big),
\end{aligned}$$

which is negative under condition $(\mathcal{H})$. Because of the form of System $(\mathcal{S}_1)$, it follows that $(\overline{F}, 0)$ is a locally asymptotically stable steady state. Finally, a standard comparison argument for Cauchy problems yields that if $0 < F(0) < \overline{F}$ and $u(\cdot) \geqslant 0$ then, we have $0 < F(t) < \overline{F}$ for all $t \geqslant 0$. □

Acknowledgements

The authors acknowledge the support of the program STIC AmSud (project 20-STIC-05) and of the Project "Analysis and simulation of optimal shapes - application to life sciences" of the Paris City Hall.

References

[1] L. Almeida, M. Duprez, Y. Privat, and N. Vauchelet. Mosquito population control strategies for fighting against arboviruses. *Mathematical Biosciences and Engineering*, 16(6):6274, 2019.

[2] L. Almeida, A. Haddon, C. Kermorvant, A. Léculier, Y. Privat, M. Strugarek, N. Vauchelet, and J. Zubelli. Optimal release of mosquitoes to control dengue transmission. *ESAIM: Proceedings and Surveys*, 67:16–29, 2020.

[3] L. Almeida, A. Léculier, and N. Vauchelet. Analysis of the "Rolling carpet" strategy to eradicate an invasive species. preprint https://hal.archives-ouvertes.fr/hal-03261142, June 2021.

[4] R. Anguelov, Y. Dumont, and I. V. Y. Djeumen. Sustainable vector/pest control using the permanent sterile insect technique. *Mathematical Methods in the Applied Sciences. Online.*, pages 1–22, 2020.

[5] R. Anguelov, Y. Dumont, and J. Lubuma. Mathematical modeling of sterile insect technology for control of *anopheles* mosquito. *Comput. Math. Appl.*, 64(3):374–389, 2012.

[6] M. S. Aronna and Y. Dumont. On nonlinear pest/vector control via the sterile insect technique: impact of residual fertility. *arXiv preprint arXiv:2005.05595*, 2020.

[7] H. Barclay and M. Mackauer. The sterile insect release method for pest control: a density-dependent model. *Environmental Entomology*, 9(6):810–817, 1980.

[8] L. Beal, D. Hill, R. Martin, and J. Hedengren. Gekko optimization suite. *Processes*, 6(8):106, 2018.

[9] P.-A. Bliman. Feedback control principles for biological control of dengue vectors. *18th European Control Conference (ECC), arXiv preprint arXiv:1903.00730*, 2019.

[10] P.-A. Bliman, D. Cardona-Salgado, Y. Dumont, and O. Vasilieva. Implementation of control strategies for sterile insect techniques. *Math. Biosci.*, 314:43–60, 2019.

[11] P. A. Bliman, D. Cardona-Salgado, Y. Dumont, and O. Vasilieva. Optimal control approach for implementation of sterile insect techniques. *arXiv preprint arXiv:1911.00034*, 2019.

[12] L. Cai, S. Ai, and J. Li. Dynamics of mosquitoes populations with different strategies for releasing sterile mosquitoes. *SIAM J. Appl. Math.*, 74(6):1786–1809, 2014.

[13] R. Cominetti and J.-P. Penot. Tangent sets to unilateral convex sets. *C. R. Acad. Sci. Paris Sér. I Math.*, 321(12):1631–1636, 1995.

[14] C. Dufourd and Y. Dumont. Impact of environmental factors on mosquito dispersal in the prospect of sterile insect technique control. *Comput. Math. Appl.*, 66(9):1695–1715, 2013.

[15] Y. Dumont and J. M. Tchuenche. Mathematical studies on the sterile insect technique for the Chikungunya disease and *aedes albopictus*. *J. Math. Biol.*, 65(5):809–854, 2012.

[16] M. Duprez, R. Hélie, Y. Privat, and N. Vauchelet. Optimization of spatial control strategies for population replacement, application to *Wolbachia*. *ESAIM, Control Optim. Calc. Var.*, 27:30, 2021. Id/No 74.

[17] V. A. Dyck, J. Hendrichs, and A. Robinson. *Sterile insect technique: principles and practice in area-wide integrated pest management*. Springer, 2006.

[18] L. Esteva and H. M. Yang. Mathematical model to assess the control of aedes aegypti mosquitoes by the sterile insect technique. *Math. Biosci.*, 198(2):132–147, 2005.

[19] K. R. Fister, M. L. McCarthy, S. F. Oppenheimer, and C. Collins. Optimal control of insects through sterile insect release and habitat modification. *Math. Biosci.*, 244(2):201–212, 2013.

[20] T. H. Gronwall. Note on the derivatives with respect to a parameter of the solutions of a system of differential equations. *Annals of Mathematics*, pages 292–296, 1919.

[21] J. Hedengren, J. Mojica, W. Cole, and T. Edgar. Apopt: Minlp solver for differential and algebraic systems with benchmark testing. In *Proceedings of the INFORMS National Meeting, Phoenix, AZ, USA*, volume 1417, page 47, 2012.

[22] M. W. Hirsch and H. Smith. Monotone dynamical systems. In *Handbook of differential equations: ordinary differential equations*, volume II, pages 239–257. Elsevier B. V., Amsterdam, 2005.

[23] M. Huang, X. Song, and J. Li. Modelling and analysis of impulsive releases of sterile mosquitoes. *Journal of biological dynamics*, 11(1):147–171, 2017.

[24] E. B. Lee and L. Markus. *Foundations of optimal control theory [by] E. B. Lee [and] L. Markus*. Wiley New York, 1967.

[25] J. Li and Z. Yuan. Modelling releases of sterile mosquitoes with different strategies. *Journal of biological dynamics*, 9(1):1–14, 2015.

[26] P. Mattila. *Geometry of sets and measures in Euclidean spaces: fractals and rectifiability*. Number 44 in Cambridge Studies in Advanced Mathematics. Cambridge university press, 1999.

[27] W. Rudin. Functional analysis mcgraw-hill inc. *London, New York*, 1991.

[28] B. Stoll, H. Bossin, H. Petit, and M. Marie, J.and Cheong Sang. Suppression of an isolated population of the mosquito vector aedes polynesiensis on the atoll of tetiaroa, french polynesia, by sustained release of wolbachia-incompatible male mosquitoes. In *Conference: ICE - XXV International Congress of Entomology, At Orlando, Florida, USA.*, 2016.

[29] M. Strugarek, H. Bossin, and Y. Dumont. On the use of the sterile insect release technique to reduce or eliminate mosquito populations. *Appl. Math. Model.*, 68:443–470, 2019.

[30] R. Thomé, H. Yang, and L. Esteva. Optimal control of *aedes aegypti* mosquitoes by the sterile insect technique and insecticide. *Math. Biosci.*, 223(1):12–23, 2010.

[31] X. Zheng et al. Incompatible and sterile insect techniques combined eliminate mosquitoes. *Nature*, 572:56–61, Aug 2019.