



Universidade do Estado do Rio de Janeiro
Centro de Tecnologia e Ciência
Instituto de Matemática e Estatística

Michel Antonio Tosin Caldas


**Modeling and uncertainty quantification in the nonlinear
dynamics of epidemiological phenomena: Application to Zika
virus and COVID-19 outbreaks**

Rio de Janeiro

2021

Michel Antonio Tosin Caldas

**Modeling and uncertainty quantification in the nonlinear dynamics of
epidemiological phenomena: Application to Zika virus and COVID-19
outbreaks**



Dissertação apresentada, como requisito parcial para obtenção do título de Mestre, ao Programa de Pós-Graduação em Ciências Computacionais, da Universidade do Estado do Rio de Janeiro.

Orientador: Prof. Dr. Americo Barbosa da Cunha Junior

Coorientador: Prof. Dr. Flávio Codeço Coelho

Rio de Janeiro

2021

CATALOGAÇÃO NA FONTE
UERJ / REDE SIRIUS / BIBLIOTECA CTC-A

C972 Caldas, Michel Antonio Tosin.
Modeling and uncertainty quantification in the nonlinear dynamics of epidemiological phenomena: application to Zika vírus and COVID-19 outbreaks/ Michel Antonio Tosin Caldas. – 2021.
118 f. : il.


Orientador: Americo Barbosa da Cunha Junior
Coorientador: Flávio Codeço Coelho
Dissertação (Mestrado em Ciências Computacionais) - Universidade do Estado do Rio de Janeiro, Instituto de Matemática e Estatística.

1. Epidemiologia - Modelos matemáticos – Teses. 2. Epidemiologia - Métodos estatísticos – Teses. 3. Zika vírus - Teses. 4. COVID-19 (Doença) - Teses. I. Cunha Junior, Americo Barbosa da. II. Coelho, Flávio Codeço. III. Universidade do Estado do Rio de Janeiro. Instituto de Matemática e Estatística. IV. Título.

CDU 616-036.22

Patricia Bello Meijinhos CRB7/5217 -Bibliotecária responsável pela elaboração da ficha catalográfica

Autorizo, apenas para fins acadêmicos e científicos, a reprodução total ou parcial desta dissertação, desde que citada a fonte


Assinatura


Data

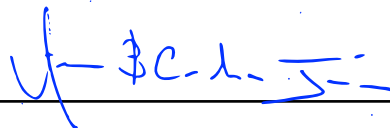
Michel Antonio Tosin Caldas

Modeling and uncertainty quantification in the nonlinear dynamics of epidemiological phenomena: Application to Zika virus and COVID-19 outbreaks

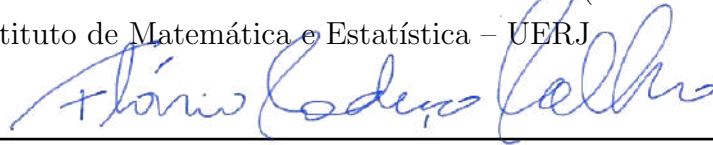
Dissertação apresentada, como requisito parcial para obtenção do título de Mestre, ao Programa de Pós-Graduação em Ciências Computacionais, da Universidade do Estado do Rio de Janeiro.

Aprovada em 06 de Agosto de 2021.

Banca Examinadora:



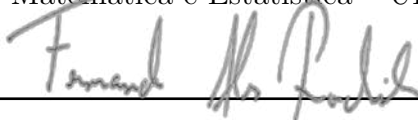
Prof. Dr. Americo Barbosa da Cunha Junior (Orientador)
Instituto de Matemática e Estatística – UERJ



Prof. Dr. Flávio Codeço Coelho (Coorientador)
Fundação Getúlio Vargas



Prof.ª Dra. Zochil González Arenas
Instituto de Matemática e Estatística – UERJ



Prof. Dr. Fernando Alves Rochinha
Universidade Federal do Rio de Janeiro



Prof. Dr. Rogério Luis Rizzi
Universidade Estadual do Oeste do Paraná

Rio de Janeiro

2021

ACKNOWLEDGMENTS

I would like to express my deepest appreciation to all those who provided me the possibility to complete this work. A master's project is always a challenge, but the pandemic scenario during the last part of the research caused me great physical and mental wounds. Several people were important during my commute and, of course, there is no way to talk about all of them here in details. Although, have the moral duty to give some special thanks.

To all my colleagues from the Americo's team with who I learned so much during the last few years. With you I improved my critical sense and my presentation skills. In particular, I highlight my friends Marcos Issa and Diego Matos with who I shared several experiences and know that I can ever count on.

I am strongly grateful to the Rio de Janeiro State Research Support Foundation (Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro – FAPERJ) for the financial support given during the first two years of the project identified by code E-26/203.046/2018, extended for another 3 months to mitigate the effects of the pandemic.

To the graduate program of computational sciences (Programa de Pós-Graduação em Ciências Computacionais – PPG-CComp) I express the great importance of the knowledge acquired in class and all the assistance during my master's course.

I need to thank and apologize to my physiotherapist Rogério Rocha for the treatment of my muscle issues. I know I was not the most cooperative patient in the beginning, but you did your best to took my physical pain away and allow me to go back to work.

To my co-advisor Dr. Flávio Coelho I appreciate all the tips and experience on the field of modeling epidemiology phenomena you divided with me.

About my adviser Dr. Americo Cunha Jr. I have to recognize all the guidance you gave me since we met in 2015. But here, I will focus on the last months. More than everything, you understood my physical and mental issues and was careful to allow me time and space needed to advance the research while respecting my limitations.

To my mother for be on my side during all the stages of the COVID-19 pandemic. I know these last months were not easy, mainly due my physical limitations. As I am a person who has difficulty in admitting your pain, I register here the importance you had during that time (not ignoring the previous years, obviously).

Finally, I dedicate this last acknowledgment to the most important person in my life: My love, Thaís. During the pandemic I have been in very bad places several times and the truth is that you are the reason that I did not give up my master's degree or worse. For that, I will never be grateful enough.

In addition, to all the others that, in some way, helped in my trajectory, but were not highlighted previously.

Two roads diverged in a wood and I - I took the one
less traveled by, and that has made all the difference.

Robert Frost

ABSTRACT

CALDAS, Michel Antonio Tosin. **Modeling and uncertainty quantification in the nonlinear dynamics of epidemiological phenomena**: application to Zika virus and COVID-19 outbreaks. 2021. 118 f. Dissertação (Mestrado em Ciências Computacionais) - Instituto de Matemática e Estatística, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2021.

Outbreaks due infectious diseases has been drawing attention of the scientific community in the last few years. The recognition of the aggressive effects created for the health and economy of the population worldwide made researchers from the most diverse areas of knowledge turn their resources into projects inside the theme. The present work presents and apply a framework for uncertainty quantification in epidemiological models. This is based on use global sensitivity analysis by Polynomial Chaos Expansion-based Sobol indices, combined with the Maximum Entropy Principle. The first allows to identify the most relevant input parameters, while the second one orients the construction of least biased distributions for those inputs. Then, a Monte Carlo simulation is executed to analyze the outcome stochastic process obtained through the model. The framework was applied in the epidemiological scenarios of Zika virus in Brazil and COVID-19 in Rio de Janeiro city, allowing to extract some important statistics about each outbreak. A compartmental model is employed in the first scenario, while the multi-waves dynamics of the second scenario is described by a Beta logistic growth model. Before riding the robustness study, calibration results are performed to put the quantities of interest obtained from theses models in a shape closer to the real data. Additional discussions are made about how to use sensitivity analysis results to update the knowledge about the parameters, and guide model selection.

Keywords: Epidemiological modeling. Nonlinear dynamics. Model calibration. Global sensitivity analysis. Uncertainty quantification.

RESUMO

CALDAS, Michel Antonio Tosin. **Modelagem e quantificação das incertezas na dinâmica não-linear de fenômenos epidemiológicos**: aplicação em surtos de Zika vírus e COVID-19. 2021. 118 f. Dissertação (Mestrado em Ciências Computacionais) - Instituto de Matemática e Estatística, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2021.

Surtos por doenças infecciosas têm tomado atenção da comunidade científica geral nos últimos anos. O reconhecimento dos agressivos efeitos gerados para a saúde e economia das populações ao redor do mundo, fez com que pesquisadores das mais diversas áreas do conhecimento voltassem seus recursos para projetos nesse tema. O presente trabalho apresenta e aplica um framework para quantificação de incertezas em modelos epidemiológicos. Este é baseado em usar análise de sensibilidade global por índices de Sobol baseados em Expansão em Polinômios Caos, combinado com o Princípio do Máximo de Entropia. O primeiro permite identificar os parâmetros de entrada mais relevantes, enquanto que o segundo orienta a construção de distribuições menos enviesadas para essas entradas. Assim, uma simulação de Monte Carlo é executada para analisar o processo estocástico de saída obtido através do modelo. O framework foi aplicado nos cenários epidemiológicos de Zika vírus no Brasil e de COVID-19 no município do Rio de Janeiro, permitindo extrair algumas estatísticas importantes sobre cada surto. Um modelo comportamental é empregado no primeiro cenário, enquanto a dinâmica multi ondas do segundo cenário é descrita por um modelo de crescimento Beta logístico. Antes de conduzir os estudos de robustez, resultados de calibração são incluídos para por as quantidades de interesse obtidas por esses modelos numa forma mais próxima dos dados reais. Discussões adicionais são feitas sobre como utilizar resultados de análise de sensibilidade para atualizar o conhecimento sobre os parâmetros, e guiar seleção de modelos.

Palavras-chave: Modelagem epidemiológica. Dinâmica não o linear. Calibração de modelos. Análise de sensibilidade global. Quantificação de incertezas.

LIST OF FIGURES

Figure 1 - Summary of the epidemiology principles for outbreak response.	15
Figure 2 - Schematic representation of the infected period.	22
Figure 3 - Contagion of COVID-19: Comparison between countries for the cumulative number of cases until 05/30/20.	26
Figure 4 - Schematic of a logistic growth cumulative curve. The growth occurs in three main regimes: (i) initial exponential growth; (ii) intermediary linear growth passing through the inflection point; (iii) saturated growth towards to the carrying capacity.	27
Figure 5 - Illustration of the effect of the parameter α for the symmetry in the model response.	28
Figure 6 - Schematic diagram for the SIR compartmental model with demographic dynamics.	31
Figure 7 - Schematic diagram for the SEIR compartmental model.	33
Figure 8 - Schematic diagram for the double population SIR-SIR model compartmental model.	34
Figure 9 - Schematic representation of model operator.	36
Figure 10 - Schematic representation of a Monte Carlo process.	40
Figure 11 - Schematic representation of the CE method for a Gaussian distribution family.	42
Figure 12 - Schematic representation of the Uncertainty Quantification framework.	48
Figure 13 - Time series for the weekly number of new and cumulative cases from Zika in Brazil in 2016.	50
Figure 14 - Schematic diagram for the SEIR-SEI compartmental model.	52
Figure 15 - Validation plots for the 2th,7th and 20th EWs, using the PCE surrogate for the QoI \mathcal{N}	54
Figure 16 - Total, First and Second orders Sobol indices for the model number of new cases.	55
Figure 17 - Comparison between first calibration (blue) and new calibration (green) obtained due to sensitivity analysis. Parameter values and IC can be found in Table 3.	56
Figure 18 - The 95%-confidence bands for the model cumulative number of cases, in Scenarios A to E.	59
Figure 19 - The 95%-confidence bands for the model number of new cases, in Scenarios A to E.	60
Figure 20 - Histogram and estimated PDF (kernel density) of the time average for the cumulative number of cases, in Scenarios A to E.	61

Figure 21 - Histogram and estimated PDF (kernel density) of the time average for the number of new cases, in Scenarios A to E.	62
Figure 22 - Marginal distributions for the random inputs, in Scenarios A to E. . . .	63
Figure 23 - Time series for the daily number of new (left) and cumulative (right) deaths from COVID-19 in Rio de Janeiro since March 17, 2020.	66
Figure 24 - First wave time series for the daily number of new (left) and cumulative (right) deaths in each wave from the first sequential calibration step. . .	69
Figure 25 - Second wave time series for the daily number of new (left) and cumulative (right) deaths in each wave from the second sequential calibration step.	71
Figure 26 - Third wave time series for the daily number of new (left) and cumulative (right) deaths in each wave from the third sequential calibration step. . .	71
Figure 27 - Fourth wave time series for the daily number of new (left) and cumulative (right) deaths in each wave from the fourth sequential calibration step.	72
Figure 28 - Full time series for the daily number of new (left) and cumulative (right) deaths in each wave obtained through the sequential calibration process.	72
Figure 29 - Validation plots for the PCE surrogate constructed for the QoI \mathcal{D}	74
Figure 30 - Total Sobol indices for the daily number of new deaths per wave. . . .	75
Figure 31 - The 95%-confidence bands for the model cumulative number of cases (left) and number of new deaths (right).	76
Figure 32 - Histogram and estimated PDF (kernel density) of the time average for the model cumulative number of cases (left) and number of new deaths (right).	77
Figure 33 - The 95%-confidence bands for the model parameters.	78

LIST OF TABLES

Table	1 - Correspondence between random variable distributions and the optimal family of orthonormal polynomials.	43
Table	2 - Lower and upper bounds of the parameters, and initial conditions. . . .	53
Table	3 - Result parameters from the first calibration (blue) and the new one (green).	56
Table	4 - 95% confidence intervals for the attack rate, peak value and peak location, in Scenarios A to E.	65
Table	5 - Lower and upper bounds for the model parameters, its estimated value, mean and standard deviation.	70
Table	6 - Estimated 95% confident intervals for the final cumulative number of deaths and total increase of deaths, for each wave.	77
Table	7 - Based SEAITRD model initial conditions and parameters supports. . .	94
Table	8 - Classification values for each candidate model.	94

LIST OF SYMBOLS

t	Time
S	Number of Susceptible
E	Number of Exposed
I	Number of Infectious
R	Number of Recovered
C	Cumulative number of infected
D	Cumulative number of deceased
N	Total population
\mathcal{N}	Number of new cases
\mathcal{D}	Number of new deaths
h	Human index
v	Vector index
w	Epidemiological week
J	Discrepancy between data and model response
\mathcal{J}	Objective function
\mathcal{M}	Model operator
\mathcal{M}^{PC}	PCE Model operator
f	Probability density function
S_u	Sobol indices
S_i^T	Total order Sobol indices
\mathbf{x}	Parameters vector
\mathbf{y}	Model observable
\mathbf{X}	Random parameters vector
\mathbf{Y}	Random model observable
β	Transmission rate
α	Exposed rate/asymmetry rate
γ	Recovery rate
r	Vector birth/mortality rate
q	Final growth rate
p	Initial growth rate

K	Carrying capacity
τ	Time of transition of waves
ρ	Transition parameter between waves
μ	Mean value
σ	Standard deviation
\mathbf{v}	Hyper parameters set
ϵ	Elite sample set
θ	Smoothing constant
ϑ	Smoothing exponent
ς	Smoothing parameter
ψ	Orthogonal polynomial bases
$\boldsymbol{\alpha}$	Polynomial multi-indices
\mathcal{A}	Multi-indices truncated set
N_s	Number of samples
g	Statistical properties
λ	Lagrange multipliers

CONTENTS

	INTRODUCTION	14
1	BASICS ON EPIDEMIOLOGY	19
1.1	Epidemics and pandemics in the human history	19
1.2	Introduction to the key notions on the field	21
1.3	Protection from and during pandemics	23
1.4	COVID-19 pandemic and its effect	23
2	MATHEMATICAL MODELS IN EPIDEMIOLOGY	25
2.1	Graphical representations	25
2.2	The logistic growth	26
2.2.1	<u>The Verhulst model</u>	27
2.2.2	<u>Some generalizations</u>	28
2.2.3	<u>Multi-waves growths</u>	29
2.3	Compartmental models	30
2.3.1	<u>Classical SIR model</u>	30
2.3.2	<u>Complementary compartments</u>	32
2.4	Another approaches	34
2.5	Model calibration and selection	35
3	PROBABILISTIC AND STATISTICAL TOOLS	37
3.1	Probabilistic elements and notation	37
3.2	Cross-entropy method	39
3.3	Polynomial chaos expansion	42
3.4	Variance based sensitivity analysis	44
3.5	Maximum Entropy Principle	46
3.6	Uncertainty quantification framework	48
4	ZIKA VIRUS OUTBREAK IN BRAZIL	50
4.1	Data set	50
4.2	Mathematical modeling	51
4.3	Sensitivity analysis	52
4.4	Calibration improvement	55
4.5	Uncertainty propagation	57

4.6	Some conclusions	64
5	COVID-19 PANDEMIC IN RIO DE JANEIRO	66
5.1	Data set	66
5.2	Mathematical modeling	67
5.3	Model calibration	68
5.4	Sensitivity analysis	69
5.5	Uncertainty propagation	76
5.6	Some conclusions	79
6	FINAL REMARKS	80
6.1	Research contributions	80
6.2	Main conclusions	80
6.3	Future directions	81
6.4	Scientific production and events	82
	REFERENCES	83
	APPENDIX A – Sensitivity analysis for model selection	91
	APPENDIX B – Supplementary material	95
	APPENDIX C – Scientific production	117

INTRODUCTION

This chapter is a preliminary presentation text about the covered problem, the particular aspects of study inside the theme and the justification for those choices. Also, the main goals are identified and the strategies used to achieve these objectives.

COVID-19 and the recent infectious diseases

Identified in 2019, a novel coronavirus is responsible for the recent pandemic of severe acute respiratory syndrome (SARS). Mutated from the original SARS coronavirus (SARS-CoV) the SARS-CoV-2 (novel SARS coronavirus) stole the spotlights when a set of outbreaks migrates to a public health emergency of international concern in January 2020 (WORLD HEALTH ORGANIZATION, 2020b). Despite all the efforts to contain the virus, the illness was responsible for 3.4M lost lives worldwide (until now, July 2021) (WORLD HEALTH ORGANIZATION, 2021b). Sequels from contagion has been documented as well (WORLD HEALTH ORGANIZATION, 2020c). The World Health Organization's COVID-19 response since 31 December 2019 has been exposed online for better understanding and transparency (WORLD HEALTH ORGANIZATION, 2020e).

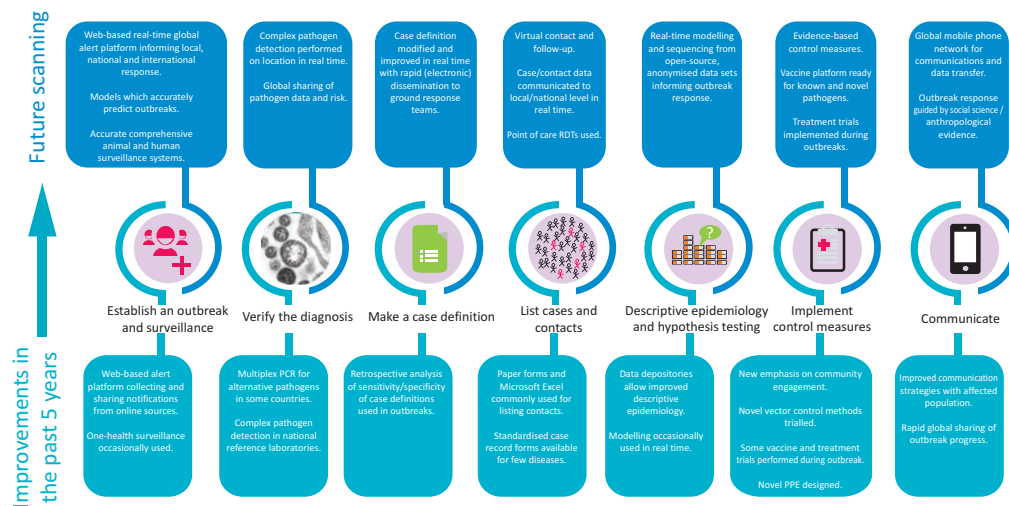
However, while the World works together to confront this thread, several other diseases are emerging or returning. Even minor lethal when compare with COVID-19, these other diseases have been effecting people resistance, cities economy and allowing subsequent illness to arrive and disseminate. In particular, most countries pass through outbreaks of arboviruses (HUANG; HIGGS; VANLANDINGHAM, 2019). This kind of comorbidity is particularly problematic to be “forgotten” on the COVID-19 pandemic because it is able to increase even during lockdowns. How testing measures has been done focusing exclusively in COVID-19, a lot of cases of arboviruses may not have been identified. Also, it is important to track the consequences that come from those. For example, the newborn microcephaly and Guillain-Barré syndrome are already associated with Zika virus infection (VALENTINE; MARQUEZ; PAMMI, 2016; DOS SANTOS et al., 2016). More details about the past and present states of the main epidemic diseases can be found in: <https://www.who.int/csr/don/en/>.

Outbreak science: Detection and response

If infectious diseases do not show signals of retreat, it is up to science to study them and understand how to combat or to prevent the infections. More precisely, Houlihan and

Whitworth (HOULIHAN; WHITWORTH, 2019), organized these goals in a set of seven epidemiology principals reunites in the Figure 1. Naturally, each one of these are covered by a different field of knowledge and therefore, the problem of infectious diseases becomes an interdisciplinary topic.

Figure 1 - Condensed summary of the epidemiology principles for outbreak response summarizing scientific progress made in the recent past and future possibilities.



Legend: PCR = polymerase chain reaction; PPE = personal protective equipment; RDT = rapid diagnostic test (should be sensitive, specific, heat-stable, cheap, simple to use, electricity-free and disposable).

Source: (HOULIHAN; WHITWORTH, 2019).

Into the elements of Figure 1, the traditional media is responsible for communication while the governments are trying to implement the control measures. These measures can be divided in preventive or combative. In the coronavirus scenario the common combative strategy is the hospitalization of the severe cases while the preventive include use of mask, social distancing and vaccination. This last one have been developed in record time and some of them were approved to be used (WORLD HEALTH ORGANIZATION, 2021a). Each country is trying to make its own agreements to obtain the largest number of vaccine options and doses as fast as possible. Even so, the immunization process is just starting. Due to the initial reduced quantity of doses available, the vaccination campaigns had be executed by steps giving preference to the vulnerable groups (WORLD HEALTH ORGANIZATION, 2020a). Again, the media participation here is crucial to explain to people why each group must be immunized first and to convince them to vaccinate (DUBÉ; VIVION; MACDONALD, 2015).

Motivation and justification

Inside outbreak situations, mathematics contributions can be brought in the sense of dynamical systems by finding good models to describe the past behavior of the studied disease and predict the future evolution from the number of new cases and deaths, beyond other relevant quantities for the spread or about the disease itself, and, by this, help to guide decision making. Nowadays several models following very distinct approaches, from differential equations to neural networks, were explored in the literature of mathematical epidemiology (WIRATSUDAKUL; SUPARIT; MODCHANG, 2018). These models, with more or less interpretability in relation with the disease dynamics in the host population, are naturally subject to several sources of errors and uncertainties due to the phenomenon being difficult by itself to be measured precisely. Thereby, by modeling also the uncertainties through a probabilistic model helps to make more robust studies about the outbreak (SOIZE, 2017; SMITH, 2014; CUNHA JR, 2017). In this discussion sensitivity analysis is also desirable to help to identify which model inputs whose the variability more affect the model response variance, and which ones are not relevant in this sense (SALTELLI et al., 2004; SALTELLI et al., 2008; TOSIN; CÔRTEZ; CUNHA JR, 2020). This kind of result is useful to simplify the probabilistic model and also reveal some nontrivial relations between the epidemic factors.

Research goals

In addition to the problem described, the general goal of this dissertation is to apply an uncertainty quantification framework, based in the use of the Polynomial Chaos Expansion-based Sobol indices together with Monte Carlo Uncertainty Propagation guided by the Maximum Entropy Principle. In particular, the framework is explored to study the distinct epidemiological scenarios of the 2016 Zika virus outbreak in Brazil and the 2020-2021 COVID-19 outbreak in Rio de Janeiro city. The idea is to approach these two scenarios using different mathematical models and strategies to cover the uncertainties. In this way, the specific research goals for each disease are as follows:

1. In the case of Zika virus:
 - (I) Describe the SEIR-SEI compartmental model used to modeling the outbreak;
 - (II) Execute a global sensitivity analysis to identify the most important model inputs during the first half of the outbreak;
 - (III) Use the previous result to update the model calibration;
 - (IV) Shows how that model performs for the most important parameters in 5 different scenarios of uncertainties;

- (V) Describe the main statistics associated with each uncertainty scenario;
2. In the case of COVID-19:
- (I) Describe a 4-waves beta logistic model for the COVID-19 deaths;
 - (II) Design a sequential calibration scheme to fit each wave of the outbreak;
 - (III) Calibrate the model response using the cross-entropy method;
 - (IV) Apply the sensitivity analysis to detect the important inputs in each wave;
 - (V) Show a scenario of small uncertainties to covers the data fluctuation;
 - (VI) Observe some statistics from the response and parameters time evolution.

Text organization

To cover the project's goals, the present dissertation is divided into 3 parts, organized in 7 chapters and 3 appendices: The first part ("Part 0") and chapter is this introduction which carries the function of describing the general points of the project. The second part is responsible for the literature review under the perspective of epidemiology, modeling approaches and statistic tools. For this, the Chapter 1 will gather an introduction on the topic of epidemiology passing by the common quantities of interest and control measures seek on this field. The mathematical base starts for the principal modeling strategies for infectious diseases, presented in the Chapter 2, combining schematic illustrations, equation representation and graphical representations to make clear how normally the quantities of interest behave depending of the set of model hypotheses. In the sequel, the Chapter 3 presents the statistical background necessary to analyze the problem in the stochastic point of view. With the main tools in hands, the Part 3.6 bring the dissertation results and conclusions. In the Chapter 4 is covered the study on the Zika scenario while the COVID-19 is analyzed in the the Chapter 5. Finally, the Chapter 6 returns the main discussion of the text and illuminates the principal results, what they represent for the problem described and some future directions that can be open from it. Additionally, the participation in events occurred in the period of this master's project and how they contributed for the whole research are described. Finally, a strategy to use Sensitivity Analysis to guide model selection is described in the Appendix A. Supplementary figures not showed in the chapters of Part 3.6 are reunited in the Appendix B, and, the visual representations from the scientific production addressed in the Chapter 6 are put together in Appendix C.

Concepts and mathematical tools

1 BASICS ON EPIDEMIOLOGY

In the last decades of the 20th century, all sorts of pathogens, including viruses like SARS-CoV-2, were favored to become dangerous and even lethal to animals and humans (PLATTO et al., 2020). The conditions for that are not in debate here, but some authors have been able to predict that a major pandemic were on its way and the destructive effect of it. Of course, these predictions are possible by analyzing the history of pandemics, which allow to apply previous successful strategies on the present issues of emergency health and avoid committing the same mistakes again. Inside this theme, this chapter has the intention of speaking about epidemiology theory by introducing the main past pandemics, the important concepts to work on the field and some ways to deal with an ongoing real situation. To finish the discussion, the final section elaborates some discussions about how the COVID-19 pandemic affects the traditional way of living.

1.1 Epidemics and pandemics in the human history

From the Plague of Athens (430–427 B.C.E) to the Novel Coronavirus pandemic (2019–), the human race suffered at the hands of contagious diseases throughout its recorded history (SCHWARTZ; KAPILA, 2021). Different in causes or transmission ways, these diseases have in common the power to destroy cities population and economy in a small time.

The first epidemic disease declared pandemic was the Plague of Justinian. Although the great wave of 541–544 remains the best documented, modern authors seem to agree on nine other episodes between the years of 557 and 700 (HAYS, 2005). Caused by the *Yersinia pestis*, Plague is a title used to name other two pandemics. The most famous is the Bubonic plague, responsible for killing more than 25 million people or at least one-third of Europe’s population during the 14th century (GLATTER; FINKELMAN, 2020), what is naturally compared with the number of deaths for COVID-19 since 2019 until now (July, 2021) (WORLD HEALTH ORGANIZATION, 2021b).

It began in 1918 one of the three pandemic influenza outbreaks occurred in the 20th century: The Spanish flu. Transmitted by the H1N1 influenza virus, the disease killed an estimated 50 million people worldwide (KAIN; FOWLER, 2019). Certainly, the spread of the disease was facilitated by the conditions generated during World War I. After that, the 1957 Asian flu (H2N2 influenza virus) and the 1968 Hong Kong flu (H3N2 influenza virus) added up 2.5 million fatalities to that account and, fifty-one years later, a novel H1N1 virus lineage (A/H1N1/2009), previously undetected in humans, started the first influenza pandemic of the 21st century (GUAN et al., 2010).

The recent history of pandemics reminds also to 2003, when severe acute respiratory syndrome (SARS) was detected in China. The failure to recognize the degree of contagion among humans at the time led the cases to grow rapidly and spread to many countries in Southeast Asia, North America, Europe, and South Africa, totaling 774 lost lives and a death rate of 9.5% when the last case was reported in that year (GUARNER, 2020).

Less than a year later, the SARS-CoV (SARS coronavirus), responsible for the SARS, mutates to a new one, consensually called MERS-CoV (Middle East respiratory syndrome coronavirus) (DE GROOT et al., 2013), putting the World on alert again. The disease caused 886 associated deaths under the 2519 laboratory-confirmed cases until January 2020 (WORLD HEALTH ORGANIZATION, 2020d). Advancing to December 2019, a new mutation of the coronavirus was found in Wuhan, China (TANG et al., 2020). The titled SARS-CoV-2 was recognized as a public health emergency of international concern in January 2020 (WORLD HEALTH ORGANIZATION, 2020b), leading many governments decreed lockdown and closing its borders (ATALAN, 2020; LAU et al., 2020). At this moment, the novel coronavirus has been reported in the whole world. Because its explosive spread, before authorities track the movement between continents, the virus reached the World and taking 3.4M lives worldwide (until now, July 2021) (WORLD HEALTH ORGANIZATION, 2021b).

Even pandemics being more notorious considering the number of people affected in different countries in the same period, many emerging epidemic diseases have created severe damage in the recent years (BLOOM; BLACK; RAPPUOLI, 2002; WATKINS, 2018). A list of each outbreak occurred in each country is described in the official vehicles of the World Health Organization (WHO) (<https://www.who.int/csr/don/en/>). Naturally, a highlight is given over the present COVID-19 pandemic. Beyond lethal illness, WHO also has especial concern with arboviruses, endemic in several regions of the World and capable of creating abrupt outbreaks. Every year, more than one billion people are infected by diseases transmitted from vectors, including malaria, Dengue, chikungunya, Chagas disease, Zika, and many more (PADMANABHAN; SESHAIYER, 2017). So its effects for the host countries cannot be neglected, even more when it is linked to the appearance of another illness (WORLD HEALTH ORGANIZATION, 2016). Considering the Brazilian scenario, the tropical weather benefits diverse arboviruses to endemic spread in the same time (DONALISIO; FREITAS; von Zuben, 2016). Unfortunately, the difficult to differentiate one from another (since the vector are the same in several situations, as well as some symptoms), delays the control.

The history of epidemics and pandemics can be sad in an anthropological sense but shows the capacity of the World of passing through crises. Probably the COVID-19 will not be the last pandemic in human history, but it shows how trust in science and union allow to create new tools to combat infectious diseases in a time never seen before.

1.2 Introduction to the key notions on the field

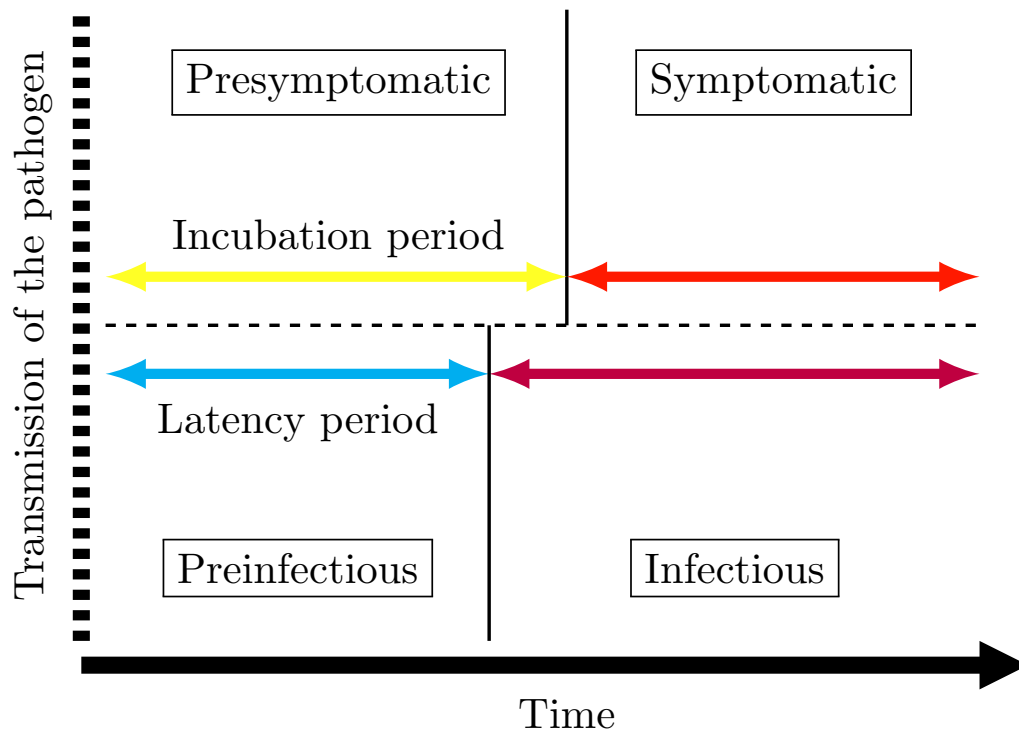
The theme of epidemiology is sensible since it often deals with the preservation of human lives. Therefore, do not confuse the basic ideas is crucial. This section lays out of demystify some known concepts of the field and present a few more.

Epidemiology is traditionally defined as the study of the distribution and determinants of health-related states or events in human populations and the application of this study to the prevention and control of health problems (SZKLO; NIETO, 2019). Nevertheless, this is a technical definition. In the practical sense, all epidemiological discussions start with the notion of a disease in the spotlight, which is caused by a pathogen (a.k.a. infectious agent), that is, any microorganism (bacterium, virus, parasite, prion) (GIESECKE, 2017). The transmission occurs when the infectious agent enters into contact with an individual capable of being infected, who is called a susceptible (or host). After that, it becomes an infected. The infected period can be analyzed on the symptoms or infectiousness perspective. From when the host was exposed to the pathogen until the onset of symptoms, it is said to be in the presymptomatic stage, or incubation (KRICKEBERG; TRONG; HANH, 2019). Thereon, the symptomatic, or clinical, stage. On this point, it is capable of being identified as infected. When the symptoms disappear, the infected period ends. On the other hand, the infected period can be also divided into preinfectious, or latent, and infectious stage depending on whether the person can already transmit the disease or not. It is completely normal to confuse these two views on the infection period (VYNNYCKY; WHITE, 2010). To clarify a little more about this issue, the Figure 2 shows an illustration of what was described above. As soon as the individual pass that, it can become recovery, disability, or death.

Infectious diseases can be transmitted by different ways. Vertical transmission is when an individual transmit to its offspring through sperm, placenta, milk, or vaginal fluids. Horizontal transmission refers to the intuitive process of an infected individual transmitting the pathogen to a susceptible contemporary (MERRILL, 2017). Additionally, transmissions can also be classified into direct, when occurs from one person to another, or indirect when the process is intermediate by an item, organism, environment or process. In epidemiology, this intermediary is named as vector (MARTCHEVA, 2015). Several diseases, as Malaria, yellow fever and Zika, are spread that way by using arthropods. Those are called arboviruses (DONALISIO; FREITAS; von Zuben, 2016).

In a broader view, when occurs cases of an illness clearly in excess of normal expectancy in a community or region, the morbidity is defined as epidemic. When restricted to a small geographical area or population, it is called outbreaks (BARRETO; TEIXEIRA; CARMO, 2006). However, if affecting or attacking the population of an extensive region, country or continent, a pandemic is characterized. Furthermore, endemic is the term used refers to the ongoing, usual, or constant presence of a disease in a community or among a

Figure 2 - Schematic representation of the infected period on the symptoms (top) and infectiousness (bottom) perspectives.



group of people. Finally, the sense of emerging infectious disease is saved to a new or an increased (as called re-emerging, to differentiate) occurrence within the last few decades (DOHERTY, 2013).

To help to understand the disease occurrence, some frequency quantities can be calculated. The idea of incidence is associated with new cases inside the host population during a given time period (SZKLO; NIETO, 2019). So, incidence times, are the times when new cases occurs, incidence proportion is the proportion of people who develop the disease during the time period, and incidence rate measures the occurrence of new cases per unit of person-time (AHRENS; PIGEOT, 2014). In a complementary way, prevalence is a quantity to the proportion of people who have the disease in that period (the cumulative portion). Finally, the idea of attack rate is a risk measure given by the final prevalence, that is, the prevalence value in the final of the time period observed.

Finally, an important concept in epidemiology is the basic reproduction number (\mathcal{R}_0). This simple number defines the average number of secondary cases that an average primary case produces in a totally susceptible population (LI, 2018). With that, it is possible to determine whether a disease will creates a epidemic or not. If $\mathcal{R}_0 < 1$, each infected host passes on the infection to fewer than one other host, and the number of new cases will decrease in time. However, when $\mathcal{R}_0 > 1$ the disease will be epidemic (BRAUER; CASTILLO-CHAVEZ; FENG, 2019).

1.3 Protection from and during pandemics

When dealing with an epidemic situation, the most important thing is to find a way to stop or, at least, slow down the contagious. For that, two types of measures are used: prevention and control. The first one is related to measures that are applied to prevent the occurrence of a disease, while the second refers to those that are applied to prevent transmission after the disease has occurred. The best example to clarify their different is the parallel between social distancing and quarantine. The first one, as the wear of masks and the personal hygiene maintaining, is about avoid to contract the disease. Differently, quarantine is about to separate sick people with a contagious disease from people who are not sick, in order to contain the spread from the disease (MERRILL, 2017).

Inside the debate of prevention, vaccination is always a “game changer”. Being able to immunize the population against the pathogen is the more effective long term way of protection. Although, development of vaccines is complex and the approval for its use must follows rigorous protocols to guarantee its efficacy and safety. Since a vaccine is authorized for studies in humans, the institution or company responsible for it should conduct those clinical trials (GIESECKE, 2017; KRÄMER; KRETZSCHMAR; KRICKEBERG, 2010; NELSON; WILLIAMS, 2014). After the efficacy’s verification, the results can be submitted to analysis from the regulatory authorities. The most important, of course, is the WHO, but the national approval is in jurisdiction of each country’ agencies. In Brazil, this is up to the *Agência Nacional de Vigilância Sanitária* (ANVISA) – or National Agency of Sanitary Vigilance, in english – , analogous to the U.S.’ Food and Drug Administration (FDA). Depending on the results (which can depend from one agency to another), the vaccine is approved to be administrated in large scale. Henceforth, the general efficacy can be observed and new side effects can be tracked.

1.4 COVID-19 pandemic and its effect

As indicated in the Section 1.1, the COVID-19 pandemic, started in 2019, has been responsible for the worst crisis in human history since the World War II. Inside this problematic, there are a lot of questions to cover and scientific advances to recognize. First of all, the first COVID-19 case in South America was sequenced in 48 hours at the Adolf Lutz Institute (DE JESUS et al., 2020), representing a historic record if compared with the mean time of 15 days needed for other international groups.

Second, the velocity of vaccine development. Although vaccination is the effective way to protect the countries population, usually it takes 15-20 years between the initial scientific discovery, vaccine licensing and policy recommendation (BLACK et al., 2020). For example, it was necessary 5 years to get a licensed vaccine for Ebola even in the public

health emergency of international concern. Nevertheless, some vaccines for COVID-19 were developed and released in about one year (WORLD HEALTH ORGANIZATION, 2021a). Of course, besides having a vaccine approved, it is important that this comes to the people. In this sense, two problems are in check: to produce the main supplies for the vaccines, fast enough to immunize the World quickly, and the logistic of transport and distribution of the product. If it is clear the impossibility of vaccination of the population, structured plans to prioritize the most vulnerable were developed. Its execution must be followed by each local health authority. At the same time, it is important to combat the anti-vaccine movement that has been gaining strength in recent years (DUBÉ; VIVION; MACDONALD, 2015).

COVID-19 pandemics also shows some weaknesses of humans. Mental issues are already one of the main concerns of the decade, and with the isolation stimulated by the pandemic, comes the sadness and fear of having clinical complications or never see some people your care about again, while there are not short-term prospects for improvement in the economic perspective. The “proliferation” of mental diseases due the pandemic quarantine are already been documented (PFEFFERBAUM; NORTH, 2020; XIONG et al., 2020) and will be an important inheritance left by the COVID-19 in the human society.

On the economic perspective, the present situation reduces with movement restriction, several business closing your doors and the unemployment rates increases all over the World (ALSAFI et al., 2020). Of course, tourist activities are stopped and international events (as Olympic Games Tokyo 2020) were postponed. A secondary effect is also the deepening of previous economic crises. In this sense, government discussions about fiscal recovery and market collaboration will be very important in the next few years.

Education is also a field affected by the isolation against the coronavirus. To avoid contact, teaching was converted to an online form (DANIEL, 2020). However, in several countries (mainly the poorest), the common families do not have financial structure to provide the necessary equipment for these home office classes, and neither the schools. Also, teachers in general were not trained to work in this condition. All these difficult the students’ learning process. The long-term effects of all those must to be observed carefully, but it is expected that the rates of learning will reduce and strategies to reinforce learning will be requested soon enough.

2 MATHEMATICAL MODELS IN EPIDEMIOLOGY

The art of creating abstractions from reality dates back to the stone age, with the painting on cave walls. Much time later, the scientific community was developing more effective representations for real objects, called mathematical models, which allow to obtain new knowledge about those phenomena by using mathematical tools. Those models are crucial on the scenario of epidemiological phenomena due the difficult of reproducing these in the laboratory. This chapter brings a brief on the state-of-the-art of mathematical modeling for epidemiology. As the topic is extensive, the idea is to cover the most common approaches used nowadays, from simplicity to complexity, focusing on differentiate them based in their hypotheses and limitations. Additionally, a discussion of model discrepancy and selection is introduced in the final section.

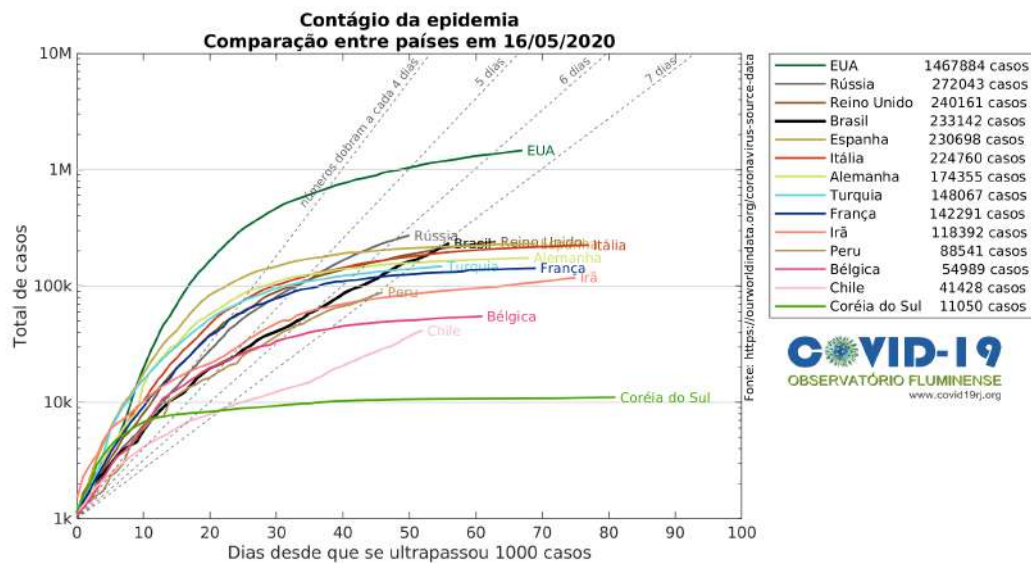
2.1 Graphical representations

The first step of any modeling process is to collect and observe data from the object of interest (BASSANEZI, 2012). Through data it is possible to see how the process is evolving during some time and extract some knowledge about it. For this proposal, graphical representations can help to more easily visualize specific aspects of the phenomenon. For epidemic scenarios, the main study object of this dissertation, the quantities of interest (QoI) can be the number of infected, recovered, dead, and others. Plot its time series helps to see the spread of the disease and how some control measures are being, or not, effective. Graphical representations are a great tool of comparison. The observation of how the disease behaves in different regions can reveal some information useful to stop its proliferation.

To illustrate how some conclusions can be obtained from graphical representation of data, the Figure 3 shows a comparison between some countries regarding contagion by COVID-19 (CUNHA JR et al., 2020). Even this graph do not consider the country population size, it can be see that some of them as South Korea, for example, seems to reach the stage where few new cases are happening. So, the disease are possible under control on this plateau. However, for others countries like Brazil, the contagion are clearly speeding up, indicating an impossibility of being contained soon. In Brazil, more restrictive control measures would be recommended. Of course, to understand better the Brazilian scenario, complementary graphical representations must be requested to explain why the national epidemic are still in a growth regime, probably not having gone through the peak of new cases yet.

In this moment, assuming a new control measure applied in Brazil, the sanitary

Figure 3 - Contagion of COVID-19: Comparison between countries for the cumulative number of cases since exceeding 1000 cases until 05/16/20.



Source: Edited from (CUNHA JR et al., 2020).

agencies desire to know how it will performs and the consequences, to be prepare to deal with those. However, graphical representation of data are not capable to predict what will happens next. Obviously, the analysis of data allows to create some intuition, but it is not sufficient, objectively, to make a prediction itself. For this, mathematical models emerge to give a relation between the characteristics of the disease and how its spreads in the studied region. Some different modeling strategies for epidemiology will be presented in the next few sections.

2.2 The logistic growth

Classically, epidemic curves are characterized by an initial exponential growth, an intermediary linear growth and a final saturated growth that converges to the final size of the epidemic, as can be see in Figure 3. It can be explained by the facility of contact new susceptible to infect in the beginning of the outbreak, reduced after long time due control measures or induced immunity. In this way, the qualitative behavior of a disease spread inside a host population is compatible to the logistic type models.

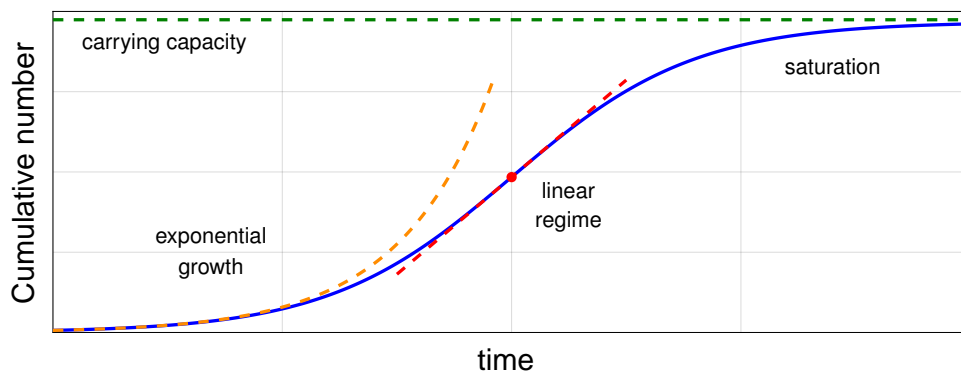
2.2.1 The Verhulst model

The logistic model was developed for Verhulst in the field of population growth (VERHULST, 1938; VERHULST, 1945; VERHULST, 1947). Influenced by the previous ideas of Malthus and Gompertz, the mathematician proposed that a population can not grow freely due to environment and resource limitations. With this assumption, its model can be written in modern notation as

$$\frac{dN}{dt} = r N \left(1 - \frac{N}{K} \right), \quad (1)$$

where the parameters r e K represent the growth rate and carrying capacity, respectively, and the cumulative population in time is expressed by $N(t)$ (BACAER, 2011). A typical qualitative form of N can be observed in the Figure 4. The inflection point indicated in red represent the change in the growth when the velocity reaches its maximum value and starts to decrease. That quantity is given by the derivative of the population equation, what in the infectious diseases scenario is the number of new cases or, locally, the measure of incidence. For practical purposes, in the epidemiological context, the cumulative number of infected will be denoted by $C(t)$ and the number of new cases by $\mathcal{N}(t)$. Similarly, $D(t)$ and $\mathcal{D}(t)$ can be used to denote the cumulative number of deaths by the studied disease, and the number of new deaths in time.

Figure 4 - Schematic of a logistic growth cumulative curve. The growth occurs in three main regimes: (i) initial exponential growth; (ii) intermediary linear growth passing through the inflection point; (iii) saturated growth towards to the carrying capacity.



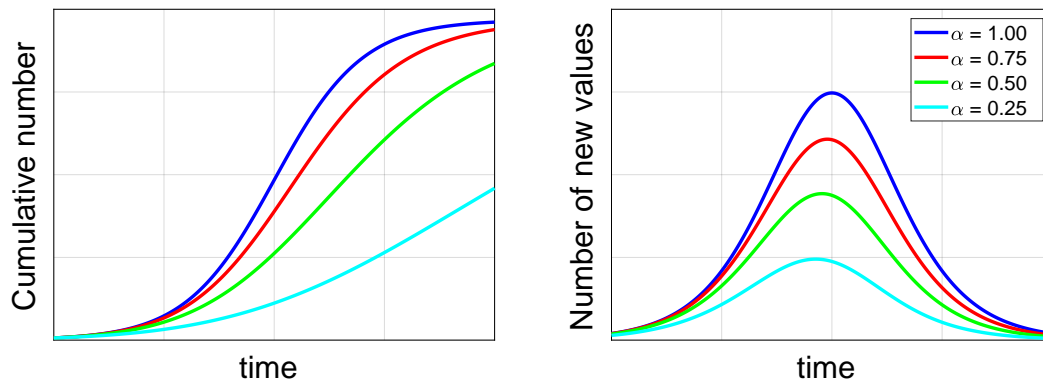
2.2.2 Some generalizations

The Verhulst model is very effective in reproducing the characteristic of the three stages of an epidemic curve, but it considers some symmetries that can not occur during an outbreak due to several reasons. In particular, the model described in the Eq. (1) implies that the inflection point (point of maximum growth or velocity, using the mechanic jargon) is half of the carrying capacity ($K/2$) and the first and third stages are symmetric, in other words, the velocity has bilateral symmetry. To contemplate the other possible scenarios, other growth models was proposed during the years. Firstly, to control the sigmoid shape of the response, Richards proposed to include an asymmetry coefficient α in the negative part of the saturation term (RICHARDS, 1959). After that addition, is formed the famous Richards growth model, given by the equation

$$\frac{dN}{dt} = r N \left[1 - \left(\frac{N}{K} \right)^\alpha \right]. \quad (2)$$

The effect of the α parameter is illustrated in the Figure 5.

Figure 5 - Illustration of the effect of the parameter α for the symmetry in the model response.



About the symmetry around the inflection point, Blumberg discussed that it is not reasonable in general, and presented a model where the initial and final growths velocities are controlled by different potency laws (BLUMBERG, 1968). By this, his model has the form

$$\frac{dN}{dt} = r N^q \left(1 - \frac{N}{K} \right)^p, \quad (3)$$

with q being the parameters which controls the initial growth regime and p commanding

how the model response approaches the carrying capacity. Note that the Blumberg model does not include the asymmetry parameter α . The reunion of the general ideas for growth models, each model and its impacts, were well documented by Tsoularis and Wallace in a review article (A.TSOULARIS; J.WALLACE, 2002), where the previous models are referred to as particular cases of the generalized logistic model described by

$$\frac{dN}{dt} = r N^q \left[1 - \left(\frac{N}{K} \right)^\alpha \right]^p . \quad (4)$$

However, a real analytic treatment for this model was only found in a recent Brazilian group work, where an implicit closed solution for the model response and critical time (time where growth regime changes) is deduced, as well as expressions for the asymptotic behaviors of the model response and velocity (VASCONCELOS et al., 2021c). In this paper, the model was named as Beta Logistic Model (BLM) for its similarities with the Beta distribution.

2.2.3 Multi-waves growths

The BLM is already sufficient for the analysis of phenomenon that evolves under a logistic type shape. Although, the model covers only one wave scenarios, that is, one set of the three states of growth. Due to interventions in the population social behavior, or environment influences, the phenomena may possibly evolves in several waves. In that case, the BLM (or its particular cases) can be extended by assigning a time dependence inside the parameters (VASCONCELOS et al., 2021a). The evolution of the parameters must be investigated in order to understand the better way to model it. For periodical scenarios, each parameter can be formulated as a sine function or a Fourier type function.

To conclude, logistic type models are relatively simple to understand and use, being well explored in several recent works on infectious diseases (VASCONCELOS et al., 2021b; CHEN; CHEN; CHEN, 2020; ZOU et al., 2020; SHEN, 2020; LIU; ZHENG; BALACHANDRAN, 2020). Nevertheless, it is important to recognize that its application for epidemiological goals is empirical (MOTULSKY; CHRISTOPOULOS, 2003). Even supposing that the growth rate r gathers the infection moment and cycle, it is unclear how the biological factors, as the latency and recovery periods, influence the growth. Some parallels can be created, but the full biological interpretability is not possible, principally in more general growth models. Another point to add is the decoupling. When the phenomenon demand the modeling of several explanatory variables, logistic growths are not able to describe how each one affect the others. A possible approach for this situation is presented on the next section.

2.3 Compartmental models

The idea of classifying people by healthy states is pretty intuitive to organize a host population and track the infection. However, it is important to remember that each individual can change its status in time. A susceptible can become an infected and an infected could migrate to recovery. By modeling the communication between the groups, also known as compartments, it is possible to understand the disease's dynamics. Such methodology is quite popular to model several recent outbreaks (DANTAS; TOSIN; CUNHA JR, 2018; ARONNA; GUGLIELMI; MOSCHEN, 2018).

2.3.1 Classical SIR model

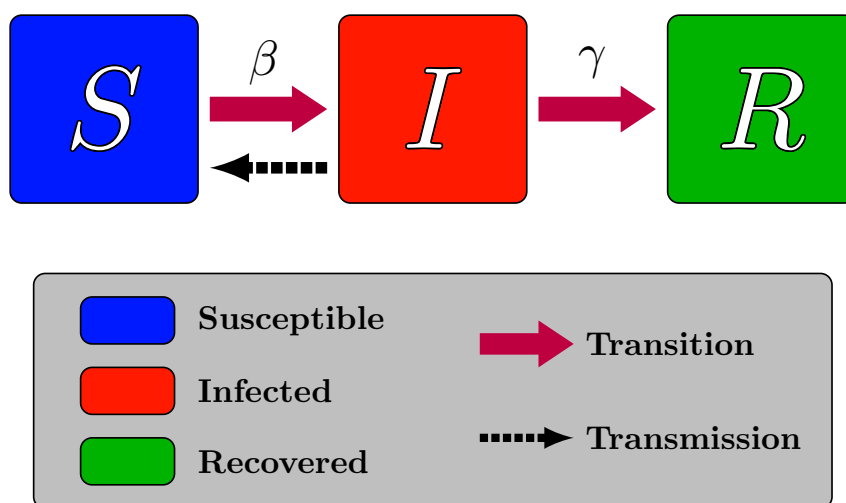
The classic work on compartmentalization strategy was presented by Kermack and McKendrick in a set of three papers (KERMACK; MCKENDRICK, 1927; KERMACK; MCKENDRICK, 1932; KERMACK; MCKENDRICK, 1933). In those, the host population was separated into three groups: The “virgins”, those who have no previous contact with the disease's pathogen; “Sick”, characterized by carrying the disease and being able to transmit it; Recovered, partially immune individuals who have recovered from at least one infection cycle. To be precise, the probability of death by the disease was also considered, but for now this hypothesis will be neglected as well as immigration, births and natural deaths. The recent texts are more familiar to call those three previous groups as Susceptible (S), Infected (I) and Recovered (R) (BRAUER; CASTILLO-CHAVEZ; FENG, 2019; BRAUER, 2017). To avoid some miss-understanding, if the recovery does not accompany immunity gain, it is better to consider the individual as a new susceptible instead of a recovered one (located in the R group). This title will be saved for those who are not more capable of developing the disease and can then be called as removed as well since there is no mortality for the pathogen in this scenario. By including the hypothesis of a homogeneously distributed population and all hosts acquiring immunity after the first infection cycle, it is obtained the most classic compartmental model found in the main mathematical epidemiology books: the SIR model (LI, 2018). Denoting the transmission rate as β , the recovery period as $1/\gamma$ and considering a constant total population size of

N , the differential formulation for the model can be written as

$$\begin{aligned}\frac{dS}{dt} &= -\beta S \frac{I}{N}, \\ \frac{dI}{dt} &= \beta S \frac{I}{N} - \gamma I, \\ \frac{dR}{dt} &= \gamma I.\end{aligned}\tag{5}$$

Here the transmission term $\beta SI/N$ is designed through frequency-dependence logic (BEGON et al., 2002), the recovery term γI contains all the processes that form the disease cycle inside the host, and the time dependence in the compartments is omitted to maintain a cleaner notation. To clarify the communication between the groups during the evolution of the disease, the Figure 6 shows a schematic diagram of the SIR model, with the transmission (by infection) highlighted in the black arrow while the transition among health states is indicated by the red ones. In addition, population dynamics can be added to the SIR model by including an input term bN in the susceptible equation to represent births and immigration, and an output term for each compartment, proportional to a ‘‘mortality rate’’, to indicate natural deaths and emigration (BRAUER; VAN DEN DRIESSCHE; WU, 2008). In that situation it is prudent to observe also the evolution of the population in time, through its own differential equation characterized by the balance of these inputs and outputs.

Figure 6 - Schematic diagram for the SIR compartmental model with demographic dynamics.



As discussed before, epidemiological data are commonly collected in the form of cumulative number of cases or number of new cases. So, it is important to extract these quantities of interest from the model. How these quantities are reported when the people

manifest symptoms, the cumulative number of cases increases when new infections occur, it can be modeled by

$$\frac{dC}{dt} = \beta S \frac{I}{N}. \quad (6)$$

The number of new cases can be obtained by taking the derivative of the cumulative number or by applying a recurrence equation as follows

$$\begin{aligned} \mathcal{N}_t &= C_t - C_{t-1}, \\ \mathcal{N}_{t_0} &= C_{t_0}. \end{aligned} \quad (7)$$

The time notation in the above equation is indicated by index because \mathcal{N} will be not necessarily in the same time unit of C . Moreover, t_0 represents the initial time of analysis.

2.3.2 Complementary compartments

The SIR framework is quite simple for didactic goals, but do not include the minor changes inside the individual during the infection cycle. For example, after the infection the individual enters in the latency period until become infectious while also enters in the incubation period until become symptomatic (See Figure 2 for more details) (KRICKEBERG; TRONG; HANH, 2019). So, it is interesting to include an intermediary compartment between the groups S and I . Right now it is important to decide if this new compartment E (for Exposed) will be left after the latency or incubation periods. Strictly speaking, it would be more correctly to use the latency period since the I compartment will now represent the infectious (capable of infect). Although, as the infectious are normally detected when the symptoms appear, they are found after the incubation period. By that, in this new SEIR compartmentalization, the transition $E \rightarrow I$ in time will be modeled under a linear rate proportional to α (given by the inverse of the incubation period). Still neglecting the demographic flow, the SEIR model, as described here, can be pictured as in the Figure 7. In addition, the system of equations to capture the dynamic behavior of

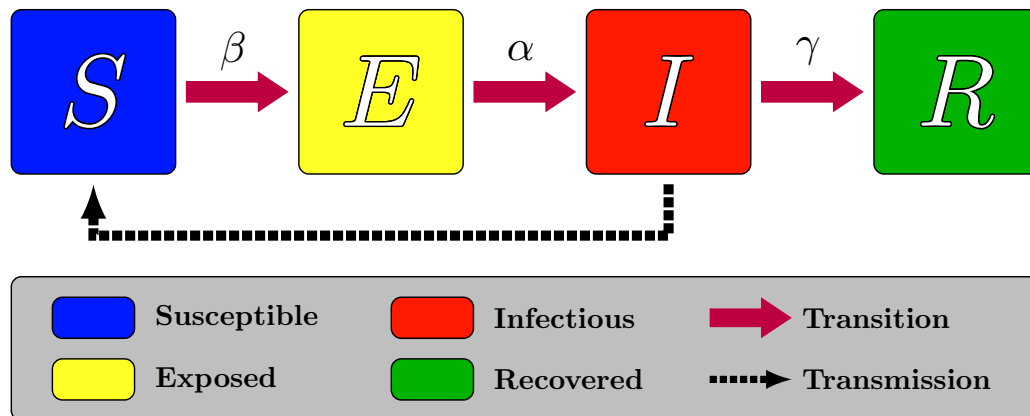
this model is given by

$$\begin{aligned}\frac{dS}{dt} &= -\beta S \frac{I}{N}, \\ \frac{dE}{dt} &= \beta S \frac{I}{N} - \alpha E, \\ \frac{dI}{dt} &= \alpha E - \gamma I, \\ \frac{dR}{dt} &= \gamma I.\end{aligned}\tag{8}$$

Again, if the cumulative number of cases changes when new infectious individuals emerge, the mathematical expression for this QoI in the SEIR model is

$$\frac{dC}{dt} = \alpha E.\tag{9}$$

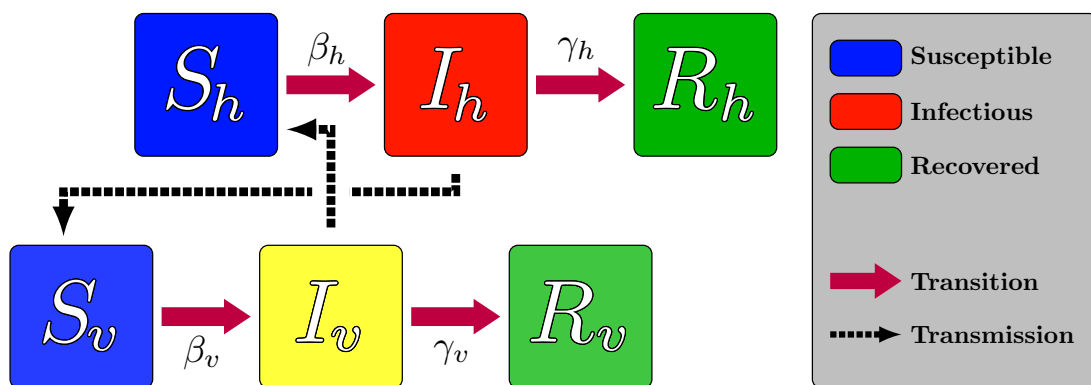
Figure 7 - Schematic diagram for the SEIR compartmental model.



Other health states can be included in order to obtain a more faithful model in the biological sense. If the disease presents a considerable portion of asymptomatic cases, split the infectious group into two compartments based in the presence or not of symptoms can be done (ARONNA; GUGLIELMI; MOSCHEN, 2018). If the individuals infected are putted into a quarantine regime, a representation for this group can be included in the model. When the number of hospitalizations affects the dynamics, this fraction of the infectious can be modeled separately, creating a new compartment. And so on. Furthermore, when dealing with two or more population susceptible to the disease, capable of cross-infection, a n -population compartmentalization can be applied (DANTAS; TOSIN; CUNHA JR, 2018; DANTAS et al., 135). The use for mosquito-borne pathogen transmis-

sion is credited to Ross and Macdonald (SMITH et al., 2012). To illustrate in a simple way, the Figure 8 shows a double population SIR model for a human-mosquito infection dynamics. The index is used to differ the human compartments from the mosquitoes (vectors) groups. The equations are analogs with the SIR model, adapting the transmission terms to the cross-form. Summarizing, compartmental models are effective to model disease states and its relations, but assumes that the population size is large enough to the mixing of members be assumed homogeneous. Also, which states are important to be considered depends on the particular characteristics of the studied disease and outbreak. To close, some text books and recent review work on the topic of compartmental models are gathered in: (LI, 2018; BRAUER; CASTILLO-CHAVEZ; FENG, 2019; BRAUER; VAN DEN DRIESSCHE; WU, 2008; SMITH et al., 2012).

Figure 8 - Schematic diagram for the double population SIR-SIR model compartmental model.



2.4 Another approaches

So far, we have considered models with homogeneous behavior and distribution. The idea in this section is to approach some ideas that come from the break of that hypotheses. To starts, the simplest way is just to consider that the disease can affect differently a host based on their age (LYRA et al., 2020). In that case, the mathematical adaptation is quite similar with an n -population compartmental model. However, instead having several populations, the total population will be divided into n indexed age sub-groups (or patches), with their respectively compartments and parameters. By assuming a model with m compartments, this metapopulation model will have a total of $m \times n$ equations (WIRATSUDAKUL; SUPARIT; MODCHANG, 2018). In a SIR compartmen-

talization example, the transmission terms will have the form $\lambda_i S_i$, where

$$\lambda_i = \sum_{j=1}^n \beta_{ij} \frac{I_j}{N_j}, \quad (10)$$

λ_i will be called the force of infection associated with the patch j and β_{ij} is the transmission rate from I_j to S_i . Similarly, it is possible to use this kind of construction in a situation where the population have different response for the pathogen based in the location (COSTA; COTA; FERREIRA, 2020). The temperature, for example, can be a factor that makes the infection more (or less) effective in an urban center than in the interior regions. So the patches will be created based on the district, or sub-region, of the territory where the host population lives. Of course, it is also possible to apply both region and age ideas to obtain a metapopulation model with double indexing, where the force of infection λ_{ij} is referred to the susceptible of the i -th age group who lives in the j -th sub-region (ROCK et al., 2014).

Still going into the heterogeneous behavior, the next step will be to consider the position as a free coordinate of the model as it is the time. So, naturally, partial differential equations become adequate to use. The common examples found in the literature explore Reaction-Diffusion or transmission Kernel (LI, 2018; KEELING; ROHANI, 2008).

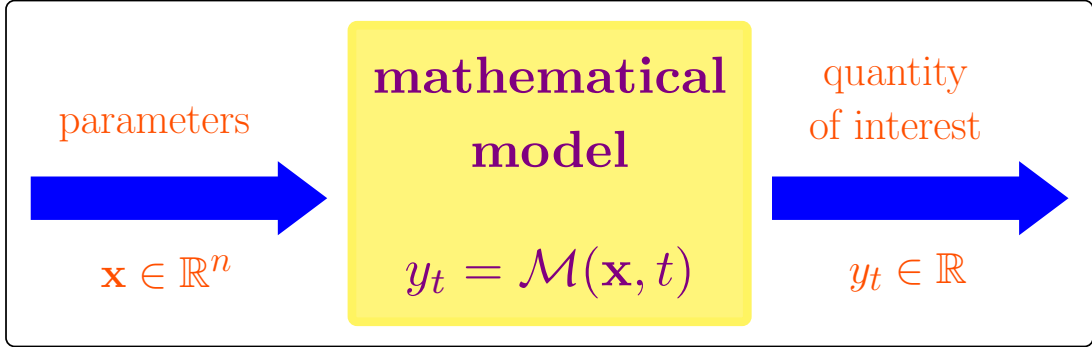
Lastly, another popular approach is the neural networks. This methodology increase with the growth of social networks in the last few years. The idea is to take into account that an individual have particular behaviors with different other individuals or groups. By modeling those distinct forms of interaction in an epidemic scenario, a more realistic spread can be elaborated. Of course, discussions about computational processing power and access to information become more important in the using of neural networks. A review work on this topic applied to epidemiology is (KEELING; EAMES, 2005).

2.5 Model calibration and selection

Before starts this section discussion, it is important to unify the notation for everything that was presented in this chapter. So, from now on, the model used in the context will be ever denoted by \mathcal{M} . The parameters vector (model inputs) will be represented as \mathbf{x} , while y will indicate the quantity of interest (QoI) of the context. With that, the Figure 9 shows how the QoI is obtained in each time instant t by the relation $y_t = \mathcal{M}(\mathbf{x}, t)$. The representation can be easily extended for m -dimensional QoI.

Assuming that the method of solution used to obtain the model observable from the inputs is verified, the quality of the model can be checked. It is extremely important to determine how faithful the model is to the represented object. For this goal, observation

Figure 9 - Schematic representation of model operator.



data is compared with the model response under the perspective of some measure of misfit J . Following the previous notation, an inverse problem can be constructed to find the input \mathbf{x}^* that satisfies

$$\mathbf{x}^* = \underset{x \in [\mathbf{lb}, \mathbf{ub}]}{\operatorname{argmin}} J(\mathbf{x}), \quad (11)$$

where \mathbf{lb} and \mathbf{ub} are, respectively, the vectors of minimum and maximum values allowed for the model inputs, and J is characterized by

$$J(\mathbf{x}) = |\mathbf{y}^{obs} - \mathbf{y}(\mathbf{x})| \quad (12)$$

where, $\mathbf{y}^{obs} = \{y_i^{obs}, i \in [1, N_t]\}$ is the dataset of N_t observations, $\mathbf{y}(\mathbf{x}) = \{\mathcal{M}(t_i, \mathbf{x}), i \in [1, N_t]\}$ is the model response and $|\cdot|$ denotes the norm adopted. More details can be found in (TARANTOLA, 2005).

Of course, when dealing with several candidate models, with similar prediction capabilities after calibration, not necessarily the closest to the data will be the best. This is even more important when the model parameters values are subjected to uncertainties. Model selection processes can be done under different criteria depending on the research objectives to choose the most suitable model (GENTLE; HÄRDLE; MORI, 2012).

3 PROBABILISTIC AND STATISTICAL TOOLS

Deterministic approaches are powerful to simulate scenarios and make predictions, among other applications. Nevertheless, the use of statistic techniques has been increased for several research fields due to its characteristic of allowing to handle with the lack of knowledge in the modeling process, parameters estimation, etc. In this chapter, some main probabilistic tools will be presented as well as some situations where they can be interesting for epidemiology objectives.

3.1 Probabilistic elements and notation

First of all, it is necessary to bring some fundamental definitions used from probability theory and the common notations found in the literature. Be a probability space characterized by the triple (Ω, \mathcal{F}, P) , where the non-empty set Ω is the sample space of all possible outcomes (or realizations, ω) for an experiment, \mathcal{F} represents the σ -algebra of the relevant events, and the probability $P : \mathcal{F} \rightarrow \mathbb{R}$ measures the likelihood of the occurrence of an event during the experiment (WASSERMAN, 2004). A random variable on it is a mapping $X : \Omega \rightarrow \mathbb{R}$ such that the image set $\text{Im}(X) = \{\omega \in \Omega : X(\omega) = x \in \mathbb{R}\} \in \mathcal{F}$, and the function that assign the probability associated to each possible value for X is called a distribution function of X (GRIMMETT; WELSH, 2017). In the sequel, X is called continuous if there exists a non-negative function f , well defined in \mathbb{R} , having the property of

$$\int_A f(x)dx = P\{X \in A\}, \forall A \subset \mathbb{R}. \quad (13)$$

This function is named the probability density function (PDF) of X and will be written $X \sim f$. On the other hand, the cumulative distribution function (CDF) of X , F , is given as follows

$$F(x) = P\{X \leq x\}, \forall x \in \mathbb{R}. \quad (14)$$

So, if the PDF exists, the mathematical relation between it and the associated CDF is

expressed by

$$\frac{dF(a)}{dt} = f(a), a \in \mathbb{R}. \quad (15)$$

Still diving into it, the CDF and PDF functions are capable of describing the random variable associated to it. To better understand the properties of X , some other measures, called moments, can also be calculated. Formally, if $X \sim f$, the n -th moment of X is constructed through the formula

$$\mathbb{E}[X^n] = \int_{-\infty}^{\infty} x^n f(x) dx. \quad (16)$$

Particularly, the first moment, or expectation, $\mathbb{E}[X]$, is very important. Also known as mean (and denoted by μ in several contexts), this quantity is useful to measure the centrality of X . The variance around the mean is given by

$$\text{var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2. \quad (17)$$

Additionally, if $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$ is a set of random variables, all the previous concepts given in this section can be generalized. With that, the relation between the n coordinates of \mathbf{X} is described by the joint CDF

$$F_{\mathbf{X}}(x_1, x_2, \dots, x_n) = P\{X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n\}, \forall \mathbf{x} \in \mathbb{R}^n, \quad (18)$$

and, over this, the concept of marginal CDFs emerge as

$$F_{X_i}(x_i) = P\{X_i \leq x_i\}, \forall x_i \in \mathbb{R}, \quad (19)$$

which will be associated with the respective marginal density functions f_{X_i} .

Besides that, if Y is a measurable function of the random variable \mathbf{X} , the properties of Y can be deduced from those of \mathbf{X} . In particular, the expectation is given by,

$$\mathbb{E}[Y(\mathbf{X})] = \int Y(\mathbf{x})f(\mathbf{x})d\mathbf{x}. \quad (20)$$

In the literature, it is very common to massively simulate the observable Y , sorting a set of realizations $\{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{N_s}\}$ from \mathbf{X} based on its statistical properties, to

numerically construct the Y statistics. This process is known as Monte Carlo method (KROESE; TAIMRE; BOTEV, 2017; CUNHA JR et al., 2014). In some situations, an auxiliary distribution g for X can be used to easily obtain an estimator for the Eq. (20). By assuming that $g(x)$ satisfies $g(\mathbf{x}) = 0 \Rightarrow Y(\mathbf{x})f(\mathbf{x}) = 0$, the Eq. (20) is rewritten as

$$\mathbb{E}_f [Y(\mathbf{X})] = \int Y(\mathbf{x})f(\mathbf{x})d\mathbf{x} = \int Y(\mathbf{x})\frac{f(\mathbf{x})g(\mathbf{x})}{g(\mathbf{x})}d\mathbf{x} = \mathbb{E}_g \left[Y(\mathbf{X})\frac{f(\mathbf{X})}{g(\mathbf{X})} \right]. \quad (21)$$

Here the index notation represents the distribution whose the expectation is related to. The distribution g will be named as an Importance Sampling density (RUBINSTEIN; KROESE, 2004). Therefore, an unbiased estimator for the expectation of Y is given by the equation

$$\hat{\mu} = \frac{1}{N_s} \sum_{i=1}^{N_s} Y(\mathbf{X}_i) W(\mathbf{X}_i), \quad (22)$$

where \mathbf{X}_i is each i th sample of $\mathbf{X} \sim g$, N_s is the number of samples and $W(x) = f(x)/g(x)$. Naturally, this estimation depends strongly on the choice of g . When $g \equiv f$, $W \equiv 1$, simplifying the estimator expression to the form

$$\hat{\mu} = \frac{1}{N_s} \sum_{i=1}^{N_s} Y(\mathbf{X}^{(i)}). \quad (23)$$

The choice of g could be also made to reduce the variance of $\hat{\mu}$ with respect to g .

Finally, regardless of the discussion about distributions, the initial section of probabilistic elements ends with a representation of the Monte Carlo method brought by the Figure 10. The notation on it is closer to the used epidemiology applications covered by the present dissertation. Thus, Y will be the stochastic process which realizations are given by the model operator \mathcal{M} , dependent of the distribution of \mathbf{X} , constructed when considering the input parameters values under the effect of uncertainties.

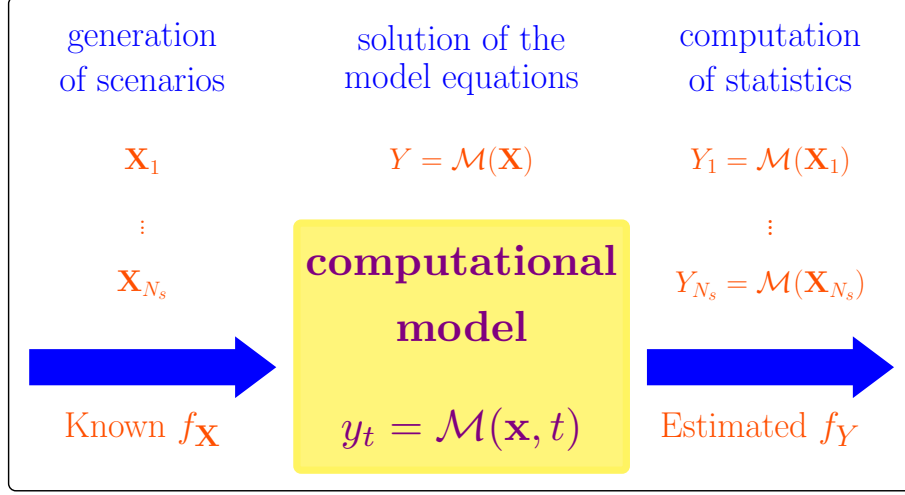
3.2 Cross-entropy method

Be the optimization problem

$$\mathbf{x}^* = \operatorname{argmax}_{x \in [\mathbf{lb}, \mathbf{ub}]} \mathcal{J}(\mathbf{x}), \quad (24)$$

for the objective function \mathcal{J} , with the assumption of unique solution to be found in the

Figure 10 - Schematic representation of a Monte Carlo process to estimate the statistics of a stochastic process $Y = \mathcal{M}(\mathbf{X})$, from the knowledge about \mathbf{X} 's distribution.



interval bounded for $[\mathbf{lb}, \mathbf{ub}]$ (TARANTOLA, 2005). Intuitively, by defining a sufficiently small region around \mathbf{x}^* and a reference level $\bar{\mathcal{J}}$, close to the maximum \mathcal{J}^* , it will be rare to find an input value inside the region that satisfies $\mathcal{J}(\mathbf{x}) \geq \bar{\mathcal{J}}$. More formally, assuming \mathbf{x} as a random variable distributed in the support $[\mathbf{lb}, \mathbf{ub}]$ by the PDF f , the probability

$$P\left\{\mathcal{J}(\mathbf{X}_i) \geq \bar{\mathcal{J}}\right\}, \quad (25)$$

will be low for all the values \mathbf{X}_i sorted through f , if $\bar{\mathcal{J}} \approx \mathcal{J}^*$. This characterizes a rare-event. The key idea of the Cross-Entropy (CE) method is to transform an maximum optimization problem into a rare-event estimation problem, where the goal is to find the distribution of importance sampling near the sampling density of theoretical greatest importance, which concentrates all its mass on the point \mathcal{J}^* (RUBINSTEIN; KROESE, 2004). For simplicity, there will be assumed that this searched distribution belongs to the family $f(\cdot; \mathbf{v})$, with its members distinguished by the hyperparameters \mathbf{v} . In this case, the problem is discover the optimum hyperparameters set \mathbf{v}^* . The common procedure is to perform a multilevel strategy. Starting from some \mathbf{v}_0 , a set of N_s samples of $\mathbf{X} \sim f(x; \mathbf{v}_0)$ will be generated. Then, the evaluation from the sample in the objective function can be done, and the results are ordered. Now, the initial level \mathcal{J}_1 is obtained by the elite set ϵ_1 of the $\rho\%$ higher values. With that, the updated hyperparameter set \mathbf{v}_1 is calculated by solving the problem

$$\max_v \frac{1}{N_\epsilon} \sum_{\mathbf{X}_k \in \epsilon_1}^{N_\epsilon} \ln f(\mathbf{X}_k; \mathbf{v}), \quad (26)$$

where N_ϵ is the size of the elite set ϵ_1 (CUNHA JR, 2021). With \mathbf{v}_1 in hands, a new set of N_s sample is sorted from which a second level characterized for the new elite set ϵ_2 can be found. Therefore, a new hyperparameter set \mathbf{v}_2 will be calculated after solving a new problem analogous to the represented in the Eq. (26). And so on. Iteratively repeating these steps, it will be generated an optimal sequence of levels and hyperparameters sets that will converge, reducing variance, to the ideal level \mathcal{J}^* and hyperparameter set \mathbf{v}^* (BOTEV et al., 2007). The process is represented for an example of Gaussian distribution f in the Figure 11. From the theoretical point of view, solving a cross-entropy-based calibration problem using PDFs from the family f is equivalent to minimize the Kullback-Leibler divergence between $f(\cdot; \mathbf{v})$ and the Dirac delta distribution $\delta(\mathbf{x} - \mathbf{x}^*)$ (KROESE; TAIMRE; BOTEV, 2017). Additionally, some adaptations can be made in the method to helps the convergence to the optimum (COSTA; JONES; KROESE, 2007). A first idea is not to fully update the t -th hyperparameter set \mathbf{v}_t in each step, but instead, to apply a linear smoothing rule

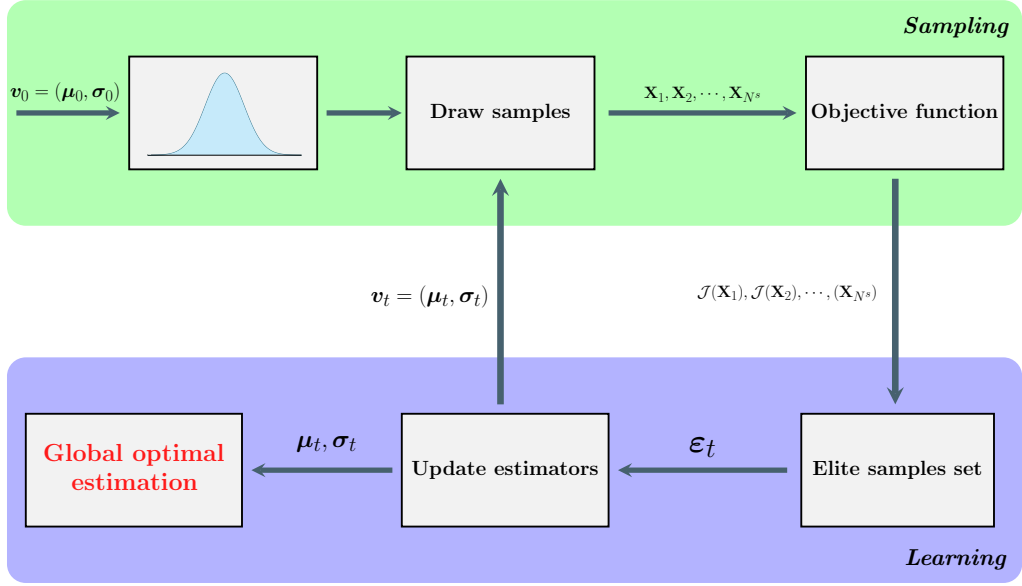
$$\mathbf{v}_t = \text{diag}(\varsigma) + (1 - \varsigma)\mathbf{v}_{t-1}. \quad (27)$$

Traditionally, the value of the smoothing parameter ς is chosen between 0.7 and 1. Although values closes to 1 have great effect in the convergence, it will reach a degenerate distribution. In that case, the method will crash in a local maximum. To avoid this, the smoothing parameter can be iteratively re-smoothed through a second rule

$$\varsigma_t = \theta + \theta \left(1 - \frac{1}{t}\right)^\vartheta. \quad (28)$$

Here ϑ is a small integer (typically between 5 and 10) and θ is a large smoothing constant (typically between 0.8 and 0.99). More details on convergence of CE methods can be found in (COSTA; JONES; KROESE, 2007). A final point to be included is that the cross-entropy method can be explored for minimum optimization problems by just using the symmetric of the original objective function ($-\mathcal{J}$) inside the method.

Figure 11 - Schematic representation of the CE method for a Gaussian distribution family.



3.3 Polynomial chaos expansion

As described before, the common way to estimate the unknown distribution f_Y , that characterizes the stochastic process $Y = \mathcal{M}(\mathbf{X}, t)$, is executing Y for a sufficient large sample of $\mathbf{X} \sim f_{\mathbf{X}}$. Naturally, the computational cost of their simulation depends of the number of samples necessary to make this estimation converges (KROESE; TAIMRE; BOTEV, 2017). Nevertheless, that also depends strongly on the complexity of performing the operator \mathcal{M} . When this last compromises the simulation time (unfeasible for high-dimensional systems), approximated models becomes very appealing to keep the running time in a region of practical use. A surrogate model for \mathcal{M} is a new operator for which the response obtained from realizations of the input \mathbf{X} is sufficiently closer to the true one, but takes much less time to be executed. In this section, it will be presented the Polynomial Chaos Expansion (PCE) metamodel for random variables.

Again, let $Y = Y(\mathbf{X}) \in \mathbb{R}$ be a random variable described as a function of a set of basic random variables, $\mathbf{X} \in \mathbb{R}^n$, with joint distribution $f_{\mathbf{X}}$. If Y also has finite variance, the functional dependence between Y and \mathbf{X} takes the form

$$Y = \mathcal{M}(\mathbf{X}) = \sum_{\alpha} y_{\alpha} \psi_{\alpha}(\mathbf{X}), \quad (29)$$

where ψ_{α} are basis of (multivariate) polynomials that are orthogonal with respect to $f_{\mathbf{X}}$,

$\boldsymbol{\alpha} \in \mathbb{N}^n$ is the multi-index used to identify the components of $\psi_{\boldsymbol{\alpha}}$, and $y_{\boldsymbol{\alpha}} \in \mathbb{R}$ are the coordinates for Y in the space characterized by the orthogonal polynomial bases (MARELLI; SUDRET, 2018; GHANEN; SPANOS, 2012). The formulation can be adapted for stochastic processes by just including a time dependence in $y_{\boldsymbol{\alpha}}$. The construction of these bases is made by tensorization

$$\psi_{\boldsymbol{\alpha}}(\mathbf{X}) = \prod_{i=1}^n \psi_{\alpha_i}(X_i), \quad (30)$$

of the univariate basis functions ψ_{α_i} , which are orthogonal with respect to f_{X_i} (PETTERSSON; IACCARINO; NORDSTRÖM, 2015; XIU; KARNIADAKIS, 2002). The Table 1 summarizes a list of the most classical polynomial families and its underlying random variables.

Table 1 - Correspondence between random variable distributions and the optimal family of orthonormal polynomials.

Distribution	Support	Orthogonal polynomials
Uniform	$[a, b]$	Legendre
Gaussian	$(-\infty, \infty)$	Hermite
Gamma	$[0, \infty)$	Laguerre
Beta	$[a, b]$	Jacobi
Poisson	$[0, 1, 2, \dots]$	Charlier

Source: (XIU, 2010).

The equality presented in Eq. (29) only occurs when all the coefficients $y_{\boldsymbol{\alpha}}$ appears in the expansion. However, the practical use of PCE metamodels comes from the truncated form

$$Y \approx \mathcal{M}^{\text{PCE}} = \sum_{\boldsymbol{\alpha} \in \mathcal{A}} y_{\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}}(\mathbf{X}), \quad (31)$$

with the truncated set of multi-indices $\mathcal{A} \subset \mathbb{R}^n$ the preserved terms are assigned. A common way to truncate a PCE is creating a sieve on the maximum polynomial degree P . Therefore, the resulted truncated set is denoted as $\mathcal{A}^{n,P} = \{\boldsymbol{\alpha} \in \mathbb{N}^n : |\boldsymbol{\alpha}| \leq P\}$, with size of

$$|\mathcal{A}^{n,P}| = \binom{n+P}{P}. \quad (32)$$

After setting the maximum polynomial degree P , a truncated PCE for Y is obtained by calculating the expansion coefficients. To that, an experimental design sample $\boldsymbol{\chi} = \{\mathbf{X}_i\}$, $i \in [1, N_s]$, should be taken by sorting $\mathbf{X} \sim f_{\mathbf{X}}$. The evaluations from that

sample, $\mathcal{Y} = \mathcal{M}(\boldsymbol{\chi})$, will be placed to feed the least-square minimization problem (TOSIN; CÔRTEZ; CUNHA JR, 2020; GHANEM; Red-Horse, 2017).

$$y_{\boldsymbol{\alpha}}^* = \operatorname{argmin} \mathbb{E} \left[\left(\mathcal{M}^{\text{PC}}(\mathbf{X}) - \mathcal{M}(\mathbf{X}) \right)^2 \right]. \quad (33)$$

Finally, after building the PCE surrogate, its accuracy must be estimated. A good error estimator is the leave-one-out (LOO) cross-validation error (ϵ_{LOO}), which is based in creating N_s surrogates $\mathcal{M}^{\text{PC}\setminus i}$ where each one is made with the exclusion of the sample point \mathbf{X}_i (BLATMAN; SUDRET, 2010). Through that, the ϵ_{LOO} value can be calculated as follows

$$\epsilon_{LOO} = \frac{\sum_{i=1}^{N_s} \left(\mathcal{M}(\mathbf{X}_i) - \mathcal{M}^{\text{PC}\setminus i}(\mathbf{X}_i) \right)^2}{\sum_{i=1}^{N_s} \left(\mathcal{M}(\mathbf{X}_i) - \hat{\mu}_Y \right)^2}, \quad (34)$$

where $\hat{\mu}_Y$ is the sample mean.

3.4 Variance based sensitivity analysis

Inside the discussion of reducing the computational cost of a Monte Carlo simulation, the dimension of the input random vector is an important factor. Possibly, a small set $\hat{\mathbf{X}}$ from the original input \mathbf{X} mainly commands the variability of the outcome Y , while the complementary set of random variables is not relevant in this sense. Even knowing the marginal input distributions and its dispersion, it is not trivial how each one really contributes for the variability of Y . To help identify those most important inputs, sensitivity analysis come into context. This section will explore the Sobol indices sensitivity analysis to select the most relevant input parameters of Y .

Given a deterministic operator $y = y(\mathbf{x})$, sensitivity analysis is basically to find a coefficient that measure the change produced into the model response due to a change applied into one of its inputs. The intuitive way to do that is by simply calculating the rate of change

$$\frac{dy}{dx_i}(\mathbf{x}_0) \approx \left| \frac{y(\mathbf{x}_\epsilon) - y(\mathbf{x}_0)}{\epsilon} \right|, \quad (35)$$

for each input coordinate x_i , in the point \mathbf{x}_0 by adding a step ϵ into x_i . Naturally, these measures are extremely local, do not considering the different behaviors that y

assumes in the domain regions that are distant from \mathbf{x}_0 . Furthermore, this strategy ignores interaction effects among input factors, i.e., effects that come from the simultaneous variation of two or more inputs in the same time. In resume, this kind of strategy will, in general, conduct to false or incomplete conclusions (SALTELLI et al., 2019). To deal with these problems, global methods has been used in recent works. The Sobol indices are a derivative-free measure for global sensitivity analysis based on the decomposition the model variance (SALTELLI et al., 2004). Thus, it is very indicated in the context of uncertainty quantification.

Returning to the random outcome $Y = \mathcal{M}(\mathbf{X})$, with X_i independent and identically distributed (iid). The Hoeffding-Sobol decomposition (SOBOL, 2001) for Y is given by

$$Y = \mathcal{M}_0 + \sum_{i=1}^n \mathcal{M}_i(X_i) + \sum_{1 \leq i < j \leq n} \mathcal{M}_{ij}(X_i, X_j) + \dots + \mathcal{M}_{1\dots n}(X_1, \dots, X_n), \quad (36)$$

where \mathcal{M}_0 is the mean value, and the terms of increasing order are conditional expectations that characterize an unique orthogonal decomposition of the model response (SALTELLI et al., 2004). They are recursively calculated as follows

$$\begin{aligned} \mathcal{M}_0 &= \mathbb{E}[Y], \\ \mathcal{M}_i(X_i) &= \mathbb{E}[Y | X_i] - \mathcal{M}_0, \\ \mathcal{M}_{ij}(X_i, X_j) &= \mathbb{E}[Y | X_i, X_j] - \mathcal{M}_0 - \mathcal{M}_i - \mathcal{M}_j, \\ &\vdots \end{aligned} \quad (37)$$

Through the Eq. (36), the model response total variance can also be decomposed as

$$\text{var}[Y] = \sum_u \text{var}[\mathcal{M}_u(X_u)] , \quad \emptyset \neq u \subset \{1, \dots, n\}. \quad (38)$$

By dividing each component of this expression for the total variance, it is obtained the set of Sobol indices, S_u , which measures the contribution of X_u for the model variance. In particular, the first order indices

$$S_i = \frac{\text{var}[\mathcal{M}_i(X_i)]}{\text{var}[Y]} \quad (39)$$

express the individual effect created due X_i (SALTELLI et al., 2008). The interaction

effect from varying the pair $\{X_i, X_j\}$ is calculated by the second order indices,

$$S_{ij} = \frac{\text{var}[\mathcal{M}_{ij}(X_i, X_j)]}{\text{var}[Y]}. \quad (40)$$

Similarly, the other order indices can be evaluated. In addition, the full contribution given by each input coordinate X_i is obtained when summing the individual and joint variance terms that include X_i , to define the total order Sobol indices (Le Gratiot; MARELLI; SUDRET, 2017)

$$S_i^T = \sum_{i \in u} S_u. \quad (41)$$

Again, those sensitivity indices can be generalized for stochastic processes by starting with a time dependence in each component of the Eq. (36).

The calculation of Sobol indices can be executed through Monte Carlo methods. In addition to the large number of simulations necessary to obtain convergence in the estimations of model total variance, it introduces cancellation errors into the calculations (KROESE; TAIMRE; BOTEV, 2017). This is even worst in dynamic systems scenarios. To avoid this problem, while also reduces the sensitivity analysis simulation cost, the PCE metamodels are called again. Due to the orthogonality properties described in the Section 3.3, the variance for a truncated PCE \mathcal{M}^{PC} for Y is given by (MARELLI; SUDRET, 2018; Le Gratiot; MARELLI; SUDRET, 2017)

$$\text{var}[\mathcal{M}^{\text{PC}}] = \sum_{\alpha \in \mathcal{A} \setminus 0} y_\alpha^2. \quad (42)$$

So, the Sobol indices can be analytically calculated as follows

$$S_u = \frac{\sum_{\alpha \in \mathcal{A}_u} y_\alpha^2}{\sum_{\alpha \in \mathcal{A} \setminus 0} y_\alpha^2}, \quad (43)$$

with $\mathcal{A}_u = \{\alpha \in \mathcal{A} : i \in u \Leftrightarrow \alpha_i \neq 0\}$ (MARELLI et al., 2018).

3.5 Maximum Entropy Principle

Until here, all the statistical ideas and methods were developed from the premise that the input PDF is well known. Some distributions are classically used based in the

support. When the model parameters are restrictively positive, for example, a Gamma distribution is preferable over a Gaussian one. However, this is an oversimplification which creates a degree of bias. Not necessarily the input distribution is compatible with the Gamma distribution. The choice of the adequate distribution in each case is not always a simple question. It is important to explore the information available about the random variables while avoiding to include sources of bias as possible. To better guide this process, this section covers the construction of maximum entropy distributions (SOIZE, 2017).

Let $\mathbf{X} = \{X_1, \dots, X_n\}$ be a set of random variables, with values inside the support $\mathcal{I}_{\mathbf{X}}$, and distributed by an unknown distribution $f_{\mathbf{X}}$. The uncertainties of \mathbf{X} can be estimated through the Shannon Entropy

$$\mathcal{E}(f_{\mathbf{X}}) = - \int_{-\infty}^{\infty} f_{\mathbf{X}} \log(f_{\mathbf{X}}) d_{\mathbf{X}} . \quad (44)$$

From it, the Maximum Entropy Principle (MaxEnt) is based on constructing the PDF of largest uncertainty from those distributions which satisfies the constraints defined by the available information (JAYNES, 1957). In mathematical terms, let a set of $m + 1$ constraints

$$\mathbb{E}[\mathbf{g}(\mathbf{X})] = \mathbf{b} , \quad (45)$$

where the real function $\mathbf{g}(\mathbf{x}) = (1, g_1(\mathbf{x}), g_2(\mathbf{x}), \dots, g_m(\mathbf{x}))$, defined in $\mathcal{I}_{\mathbf{X}}$, enforces the statistical properties $\mathbf{b} \in \mathbb{R}^{m+1}$, including the PDF normalization condition (SOIZE, 2017). The MaxEnt PDF is extracted by maximizing the Eq. (44) over the constraints. The general solution of this optimization problem has the form

$$f_{\mathbf{X}}(\mathbf{x}) = \mathbb{1}_{\mathcal{I}_{\mathbf{X}}} \exp(-\langle \boldsymbol{\lambda}, \mathbf{g}(\mathbf{x}) \rangle) . \quad (46)$$

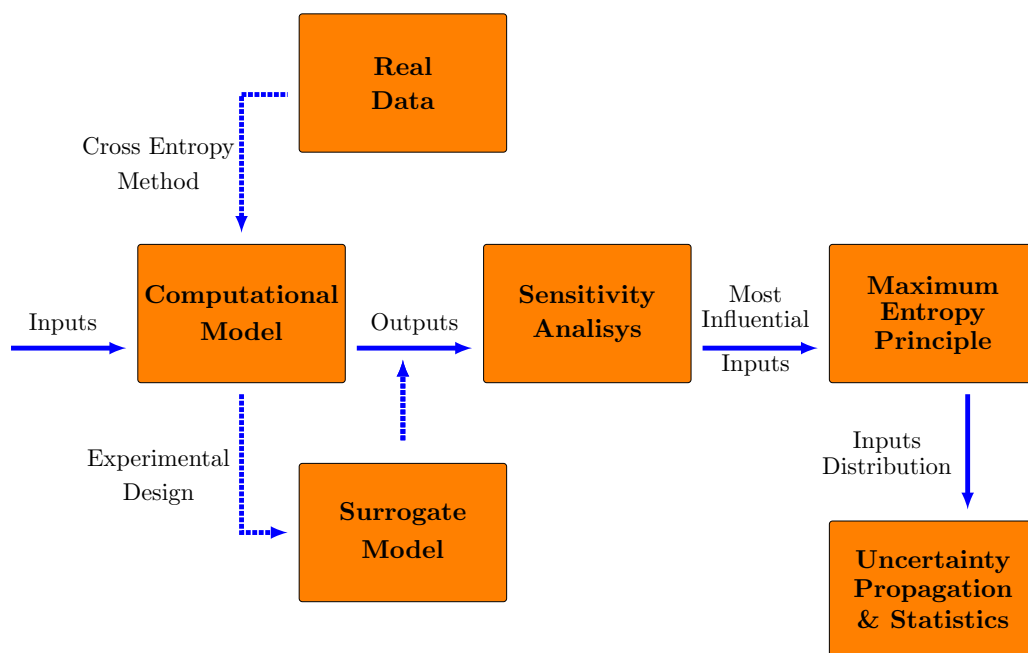
Here, $\boldsymbol{\lambda} \in \mathbb{R}^{m+1}$ indicates the vector of Lagrange multipliers, $\langle \cdot, \cdot \rangle$ is the dot product, and $\mathbb{1}_{\mathcal{I}_{\mathbf{X}}}$ represents the indicator function for the support of \mathbf{X} (SOIZE, 2017).

About particular scenarios on MaxEnt, if there is no known cross-moments, the previous construction returns independent distributions. Furthermore, when the supplied statistical information is composed by moments, the computation of the Lagrange multipliers is simplified due to the formation of a Hankel matrix (SMITH; ERICKSON; NEUDORFER, 1991). On it, when only the first moment is provided, the resulting Max-Ent PDF is a truncated exponential. If the second moment is also available, the two possible solutions are truncated Gaussians, when $0 < \lambda_3 \in \boldsymbol{\lambda} = \{\lambda_1, \lambda_2, \lambda_3\}$, or, if $\lambda_3 < 0$, a truncated distribution of the exponential family but with a quadratic exponent is found (UDWADIA, 1989).

3.6 Uncertainty quantification framework

In this chapter, some statistical tools were presented and detailed. To bring a closure in the theoretical part of the dissertation, this section put together the statistical tools in a well organized uncertainty quantification framework to be used in the Results part of this project. The idea is to use the Polynomial Chaos Expansion-based Sobol indices together with Monte Carlo Uncertainty Propagation guided by the Maximum Entropy Principle. Thus, the framework is composed by five steps, represented in the Figure 12: First off all, assuming the previous choice of a mathematical model to work with, real data from the quantity of interest will be explored to adjust some model characteristics, as initial conditions or some shape parameters, via calibration process. After that, this fitted model will be assigned as the computational model; Its input parameters will be considered as random and a experimental design will be sampled to guide the construction of a polynomial chaos expansion surrogate model; This sample will be used to perform a Sobol indices global sensitivity analysis in order to identify the most relevant input parameters; The MaxEnt principle is then applied to construct informative distributions for the relevant inputs; With this, a Monte Carlo uncertainty propagation is executed to observe the main statistics for the QoI. It is important to make it clear that the surrogate model is used only in the sensitivity analysis step. In the UQ step, the original computational model is applied. The Figure 12 also reveals that the framework can easily be adapted to others surrogate, sensitivity analysis and calibration strategies.

Figure 12 - Schematic representation of the Uncertainty Quantification framework.



Results and discussion

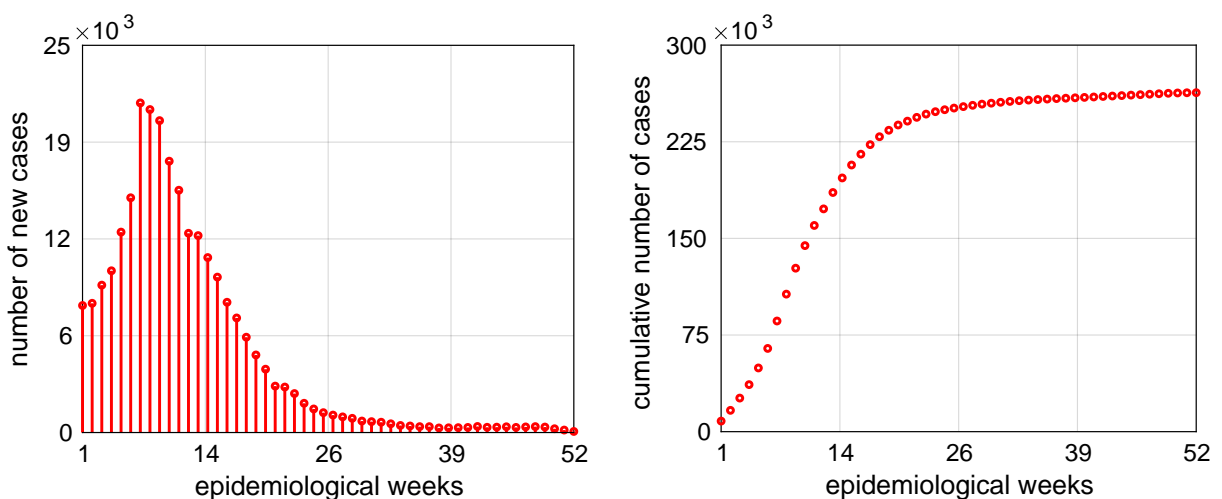
4 ZIKA VIRUS OUTBREAK IN BRAZIL

In 2016, the WHO declared the Zika virus epidemic as a global medical emergency problem due to several outbreaks spread around the globe and the increase of cases of Microcephaly and Guillain-Barré syndrome (WORLD HEALTH ORGANIZATION, 2016). Recognizing the massive reduction on the number of new cases and the apparent induced immunity to the disease, the status was removed in 2017 maintaining the preventive measures. During that period, the outbreak of Zika fever in Brazil drew a particular attention due to the sportive events that occurred in the country and the relation with the spread of the disease. This chapter approaches an uncertainty quantification study on the Brazilian outbreak of Zika virus.

4.1 Data set

The data explored here will be the number of new cases of Zika virus occurred in Brazil during the 52 epidemiological weeks (EW) of 2016 (Figure 13). For reference of the long term growth, the cumulative number of cases in each week are also represented. The information was obtained through a request send to SINAN (Sistema de Informação de Agravos de Notificação, or Notification Diseases Information System in english), and is currently available in the ZikaVD data sets (TOSIN; DANTAS; CUNHA JR., 2021) for public use.

Figure 13 - Time series for the weekly number of new (left) and cumulative (right) cases from Zika in Brazil in 2016.



4.2 Mathematical modeling

For modeling this epidemiological scenario, it will be used a double population compartmental model, which was already explored in a previous calibration study (DANTAS; TOSIN; CUNHA JR, 2018). The model is structured as follows: The two populations (humans and *Aedes aegypti* mosquitoes) are divided into a SEIR framework, removing the recovered vectors group by the consideration of a small lifetime. Including also the cumulative number of infections as a variable of interest of the system, the evolution of the Zika dynamics can be described by the set of equations

$$\begin{aligned}
 \frac{dS_h}{dt} &= -\beta_h S_h \frac{I_v}{N_v}, \\
 \frac{dE_h}{dt} &= \beta_h S_h \frac{I_v}{N_v} - \alpha_h E_h, \\
 \frac{dI_h}{dt} &= \alpha_h E_h - \gamma I_h, \\
 \frac{dR_h}{dt} &= \gamma I_h, \\
 \frac{dS_v}{dt} &= \delta N_v - \beta_v S_v \frac{I_h}{N_h} - \delta S_v, \\
 \frac{dE_v}{dt} &= \beta_v S_v \frac{I_h}{N_h} - (\alpha_v + \delta_v) E_v, \\
 \frac{dI_v}{dt} &= \alpha_v E_v - \delta I_v, \\
 \frac{dC}{dt} &= \alpha_h E_h,
 \end{aligned} \tag{47}$$

Here, the h and v indices are employed to indicate the humans and vectors (mosquitoes), respectively. The mosquitoes compartments are treated by proportions ($N_v = 1$) due to the difficulty to determine its population size. Even so, as those insects rapidly reproduce, a same birth/mortality rate δ is included to manifest this characteristic without changing the population size. The same approach was not applied to humans, measured in number of individuals and assumed to have a constant total population of $N_h = 206 \times 10^6$. The parameters β_i represent the cross-transmission rate to the infectious group I_i . The quantities $1/\alpha$ and $1/\gamma$ are the incubation and recovery period, respectively. The model equations and characteristics can be better understood on the associated diagram from the Figure 14. The two quantities of interest for the model are cumulative number of

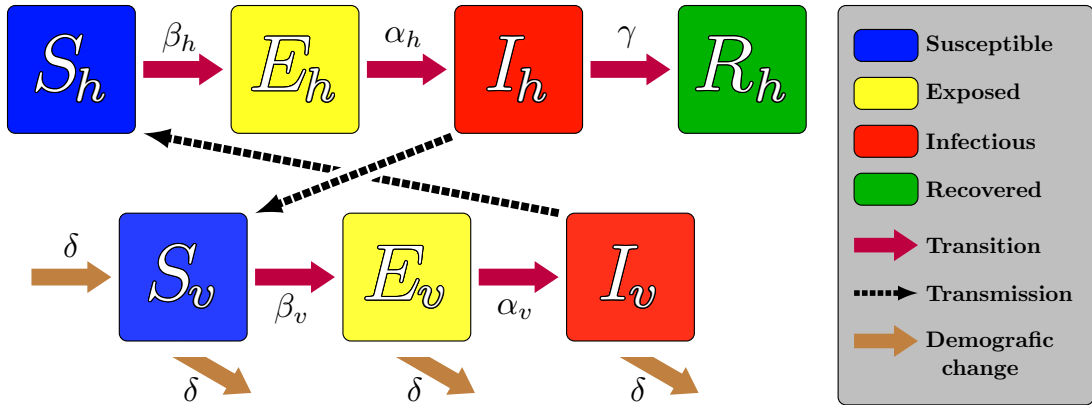
cases and the weekly number of new cases, given by

$$C(t) = \int_{t_0}^t \alpha_h E_h, \quad (48)$$

$$\mathcal{N} = C_w - C_{w-1}, \quad \mathcal{N}_1 = C_1, \quad w = 2, \dots, 52,$$

where C_w is the value of C in the w -th EW. Here the weeks are accounted in its final day, so $t = 7$ is equivalent to $w = 1$ and so on. The computational model is setting by taking a set of input parameters $(\beta_h, \alpha_h, \gamma, \beta_v, \alpha_v, \delta)$. The integration of the ordinary differential system in Eq. (47) is performed with a Runge-Kutta (4,5) method through the time interval $t = [7, 364]$, using a set of initial conditions that conserve the total of N_h and N_v in the two respective populations. Then, the QoIs are extracted for the 52 EW of the year of 2016 through the Eq. (48).

Figure 14 - Schematic diagram for the SEIR-SEI compartmental model.



4.3 Sensitivity analysis

Starting from the previous calibration study described in (DANTAS; TOSIN; CUNHA JR, 2018), the initial conditions (IC) are considered already adjusted and the parameters supports are well documented. Both are detailed in the Table 2. With that, the application of the UQ framework begins by characterizing the random inputs for the sensitivity analysis. In according with what was described in the previous section, and by the assumption of independent parameters defined in the supports presented in the Table 2 (DANTAS; TOSIN; CUNHA JR, 2018) as the unique information available to be used until here, the random variables will be represented through the random vector $\mathbf{X} \sim \mathcal{U}(\mathbf{x}; \mathbf{lb}, \mathbf{ub})$.

Table 2 - Lower and upper bounds of the parameters, and initial conditions.

parameter	β_h	α_h	γ	β_v	α_v	δ		
lb	1/16.3	1/12	1/8.8	1/11.6	1/10	1/21		
ub	1/8	1/3	1/3	1/6.2	1/5	1/11		
group	S_h	E_h	I_h	R_h	S_v	E_v	I_v	C
IC	205,953,959	6,827	10,000	29,639	0.999586	4.14×10^{-4}	0	8,201

Legend: The unit is days⁻¹ for the parameters. Human IC are counted in individuals, vector IC are in proportion.

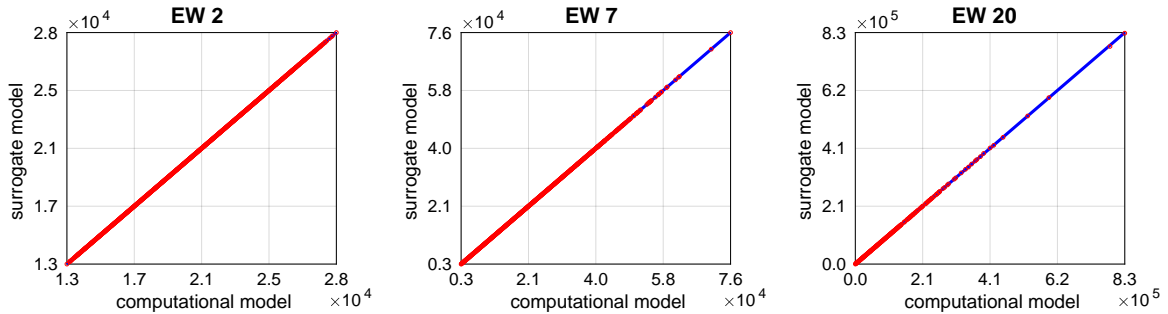
Source: References for the bounds can be found in (DANTAS; TOSIN; CUNHA JR, 2018).

Since the outbreak peak occurred in the 7th EW and until the 20th EW the most part of cases already happened, the global sensitivity analysis will be executed in the first 20 EWs only. The first week will be neglected because it is relative to the initial condition, so there is no variance in that initial point (in this study). Therefore, the final time window of EWs to be explored in the sensitivity analysis is $\{2, 3, \dots, 20\}$. The QoI used to guide the study is the number of new cases (\mathcal{N}_w). All the simulations were conducted with the fixed initial conditions listed in the Table 2. To obtain the Sobol indices, a 10 maximum degree PCE was constructed using a 20,000 sample experimental design set to obtain the 8,008 coefficients of the expansion. The relatively large sample set is applied to guarantee a precise surrogate. The leave-one-out cross-validation errors estimated in the PCE construction were of the order of 10^{-5} or less. The smaller was found in the EW 2, with value of 2.44×10^{-10} , and the higher was 1.18×10^{-5} , related to the final week of this analysis (EW 20).

To observe the approximation accuracy, a validation set of 10,000 sample was generated. Validation plots were created by comparing the true model response with the PCE surrogate response, using the identity line as a reference of dispersion between those two responses. The Figure 15 put together the validation results in the epidemiological weeks 2, 7 and 20. By doing that, it can be better noted how the PCE surrogate response is closer to the true response for the original model. The other validation curves can be found in the Appendix B.

With a trustworthy surrogate in hands, the Sobol indices could be calculated for all the possible orders. The total order sensitivity analysis result is presented in the Figure 16, together with the first, second orders indices. The other orders are omitted here because its contribution is not so relevant inside the analysis (See Appendix B). For the cases with many indices, only the 10 most higher sets are displayed. Also, the peak week was highlighted for being the most important time instant of the outbreak.

Figure 15 - Validation plots for the 2th,7th and 20th EWs, using the PCE surrogate for the weekly number of new cases.

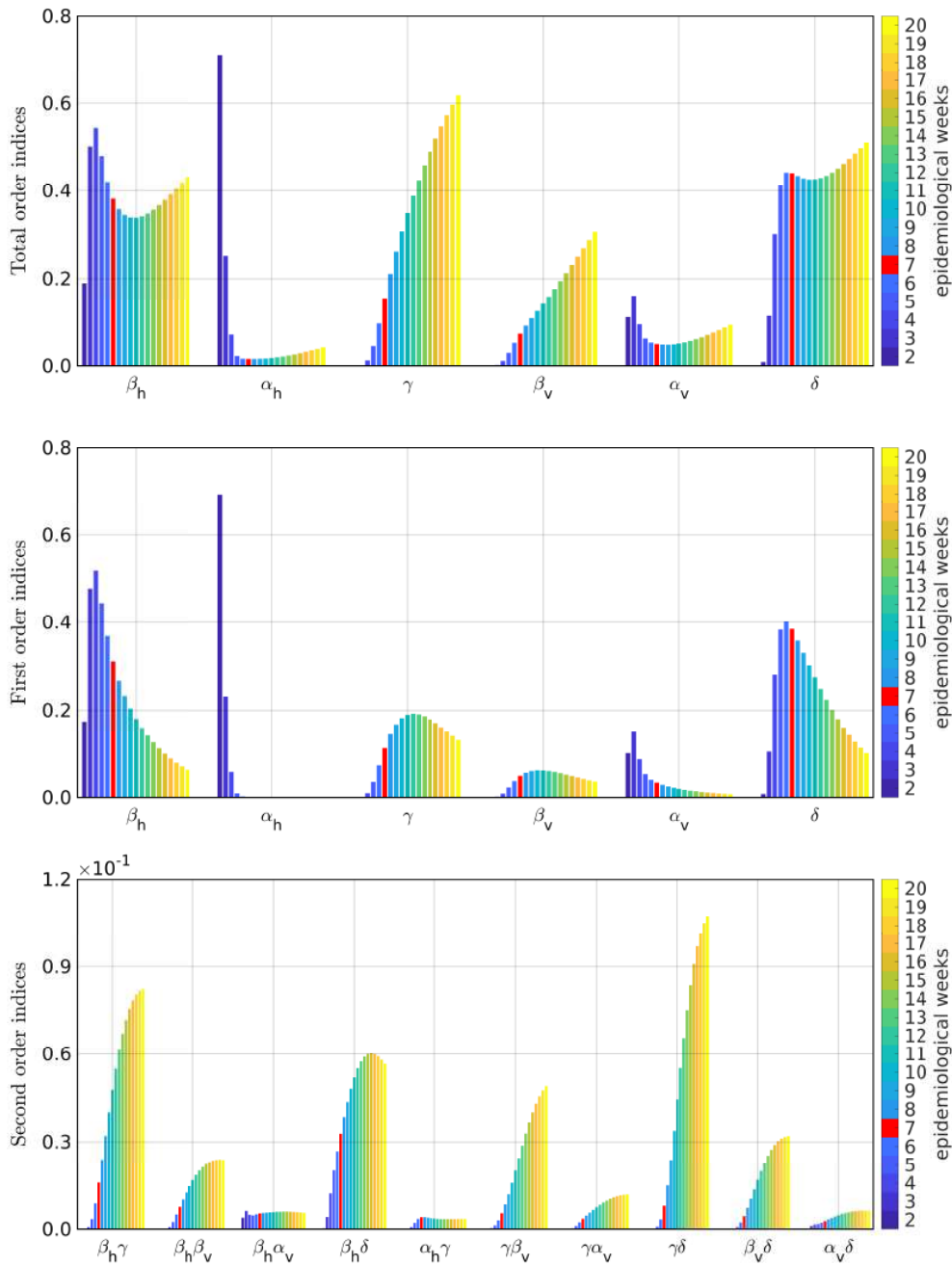


Legend: The red dots represent the sample and the blue line is the identity line.

With focus on the first order indices only, the most relevant parameters are β_h , δ and γ , with this last much less effective than the other two. The effect of the transmission rate increases in the first few weeks, because the outbreak is getting started and too much susceptible humans are available to be infected. Even so, this effect rapidly reduces, similar to what happens to α_v . The contribution given by δ increases until close to the peak week, decreasing after that. This results is not so intuitive, but occurs because δ is responsible for renew the vectors population and the new born mosquitoes will infect new susceptible humans. On the other hand, the parameters also removes vectors, reducing the transmission capacity of the vectors group. So, δ will “helps” the spread in the initial weeks and helps to contain the outbreak in the final weeks. Finally, the variability of γ and β_v have an Gaussian shape around the 11th and 12th weeks. Similar with what happens with δ , those two parameters affect the equation of new cases indirectly.

The qualitative behavior of the Total Sobol indices changes notably when including the other orders. By observing the Figure 16, most of the results of second order grows during the passing of weeks. With that, the total orders indices gain a great contribution after the peak week. The first order indices are more relevant before that. Through the total indices, the parameter γ become more competitive with β_h and δ . However, if using the 7th week as a point of reference, γ is disadvantaged. In the 7 first weeks, β_h and δ clearly are more effective that γ . This last assume the lead only when the outbreak is ending. Summarizing, the global sensitivity analysis reveals the importance of consider the cross effect between parameters, and returns the transmission rate β_h and the demographic factor δ as the most important parameters during the 20 first weeks of the outbreak.

Figure 16 - Total, First and Second orders Sobol indices for the model number of new cases.



4.4 Calibration improvement

The previous sensitivity analysis was obtained to guide the uncertainty propagation study to be executed in the next section. Although, another idea is to also use that knowledge to improve a calibration result, by focusing in the parameters that are more important in the more desirable regions. For the Brazilian Zika scenario, the most important moment of the outbreak is the peak. The blue curves of the Figure 17 shows

the model response for a calibration results obtained with the same model described in the Section 4.2, on the same conditions presented in the Table 2 (DANTAS; TOSIN; CUNHA JR, 2018). The result clearly underestimate the peak value. To improve this, a new tuning is made by only updating the pair $\{\beta_h, \delta\}$ (most important parameters). The calibration method used was the same Trust-Region-Reflective detailed in (DANTAS; TOSIN; CUNHA JR, 2018), to ensure that the gain does not come from the change of methods. The fitted parameters are detailed in the Table 3. As desired, the new result (green curves of the Figure 17) creates an increase in the values around the peak week while do not compromise the fitting in the other EWs. In the cumulative curve, the way the improved response follows the data around the peak is more clear, but, as before, disperses after that. To finish, it is important to noted how the improved tuning do not correct the response peak position, which still remains happening before the 7th EW.

Figure 17 - Comparison between first calibration (blue) and new calibration (green) obtained due to sensitivity analysis. Parameter values and IC can be found in Table 3.

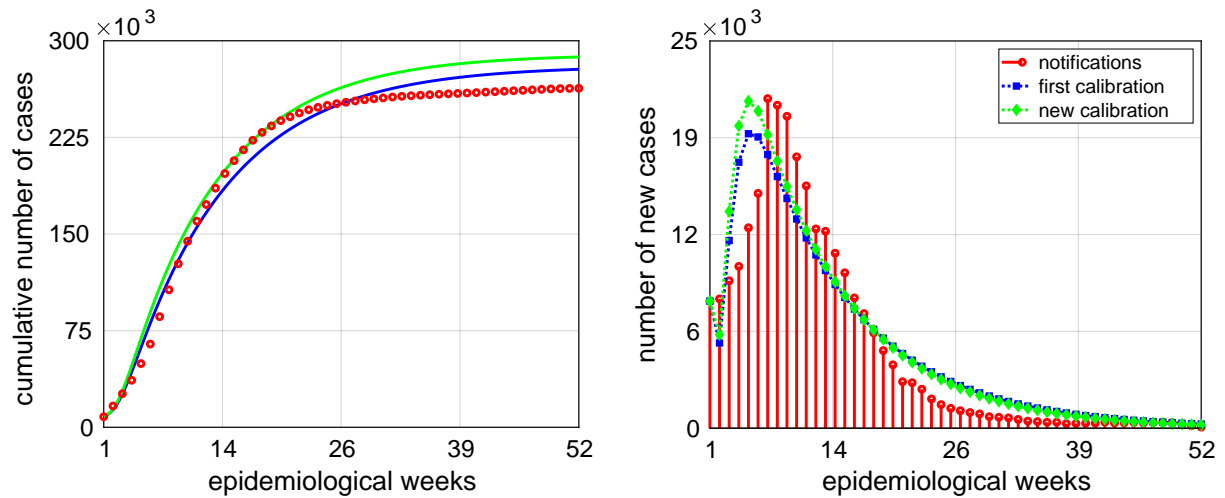


Table 3 - Result parameters from the first calibration (blue) (DANTAS; TOSIN; CUNHA JR, 2018) and the new one (green).

parameter	β_h	α_h	γ	β_v	α_v	δ
first	1/10.4	1/12	1/3	1/7.77	1/10	1/21
new	1/8.33	1/12	1/3	1/7.77	1/10	1/18

4.5 Uncertainty propagation

The new calibration guided for sensitivity analysis results was successful to correct the peak value but not its position. This maybe reveals a deficit on the used model. Also, the calibration result is an additional information available about the model parameters. According with the uncertainty quantification framework followed by this dissertation, this section covers a set of scenarios of uncertainties for the parameters selected by the sensitivity analysis executed in the Section 4.3. The idea is to investigate the model robustness to reproduce the data when uncertainties are included. The maximum entropy principle will be used to guide the construction of the inputs distributions based in the information available in each case. Based in the conclusions from the Section 4.3, the random input will be composed only by the pair $\{\beta_h, \delta\}$. Finally, the five different scenarios to be covered in this section will be the follows:

$$\left\{ \begin{array}{l} \text{random variables: } \beta_h, \delta, \\ \text{information: support, mean,} \end{array} \right. \quad (\text{Scenario A})$$

$$\left\{ \begin{array}{l} \text{random variables: } \beta_h, \delta, \\ \text{information: support, mean, } CV_{\beta_h} = 5\% , CV_{\delta} = 5\% , \end{array} \right. \quad (\text{Scenario B})$$

$$\left\{ \begin{array}{l} \text{random variables: } \beta_h, \delta, \\ \text{information: support, mean, } CV_{\beta_h} = 10\% , CV_{\delta} = 5\% , \end{array} \right. \quad (\text{Scenario C})$$

$$\left\{ \begin{array}{l} \text{random variables: } \beta_h, \delta, \\ \text{information: support, mean, } CV_{\beta_h} = 10\% , CV_{\delta} = 10\% , \end{array} \right. \quad (\text{Scenario D})$$

$$\left\{ \begin{array}{l} \text{random variables: } \beta_h, \delta, CV_{\beta_h}, CV_{\delta}, \\ \text{information: support, mean, } CV_{\beta_h}, CV_{\delta} \sim \mathcal{U}(5\%, 10\%). \end{array} \right. \quad (\text{Scenario E})$$

In these, the supports are again those from the Table 2, while the mean referred value is the result obtained in the new calibration from the Section 4.4. The variable CV represent the coefficient of variation (ratio between standard deviation and mean) around the mean, in order to introduce dispersion. Since no previous information is known about the dispersion, the assumed values are used to investigate the behavior and sensibility under this hyperparametric study for the most unbiased possible PDFs. The other model parameters remain constants with the first calibrated values (Table 3) and the initial conditions are the same used until here (Table 2). In Scenarios A to D, MaxEnt PDFs

are constructed for the random variables β_h and δ , one for each, and parameter values are sampled from these two distributions. In Scenario E, however, one sample means the following steps: a pair $(CV_{\beta_h}, CV_{\delta})$ is generated and, through that, two maximum entropy PDFs are constructed for the stochastic β_h and δ (one for each), obtaining a pair (β_h, δ) from these PDFs. The samples are generated from the MaxEnt PDFs by use of the Inverse-Transform Method (KROESE; TAIMRE; BOTEV, 2017).

The 95%-confidence bands for Scenarios A to E, respectively, are presented in the Figures 18 and 19. These were constructed by using 1,024 sample to feed a Monte Carlo Method. The convergence metric used was the sum of the second statistical moment from each compartment present in the SEIR-SEI model equations. The figures shows the convergence results can be found in the Appendix B.

The analysis of the confidence bands reveal qualitatively the influence of the input PDFs on the QoI distributions, specially the response behavior around the maximum value of new cases. In particular, the scenario A, more conservative due do not consider the mean, generated larger bands which are probably unrealistic. The other scenarios contains less data inside them, but are more informative. In general, the same percent of dispersion values was more significant when applied on the parameter δ . Also, the two random parameters seem to have little, or non, effect over the first few EW when distributed accordingly to MaxEnt PDFs. Thus, the stochastic system response fails to encompass the epidemic curve in this early stage, overpredicting. Nevertheless, the model appears to maintain its general pattern, demonstrating a robustness in the presence of uncertainties for their most sensible parameters, and following the data from the main interest region until the outbreak's end.

Several statistics can be obtained from the uncertainty quantification procedure employed in the proposed framework. The histograms and estimated PDFs of the time average for the cumulative number of cases, in all the five scenarios, are displayed in Figure 20. These results help to observe the average interval of values outbreak assumed during the outbreak, and its corresponding distribution. Even though the mean and deviation interval seem to be approximately stable in all cases analyzed, the skewness of the distribution inverts and the tail concavity changes significantly. Matching with it, each respective histogram of time average for the number of new cases is showed in the Figure 21, as well as the estimated PDFs.

The marginal distributions for the random input variables are represented in the Figure 22, where the Monte Carlo sample used in each uncertainty scenario is also plotted. The parameter β_h presented exponential behavior in the first four scenarios and Gaussian distribution on the last. The mass of probability is concentrated in values lower or equal than the mean given as information. On the other hand, δ only showed exponential distribution in the first scenario. Thus for the other four, the parameter has high possibilities to assume values bigger than the mean (inside the support).

Figure 18 - The 95%-confidence bands for the model cumulative number of cases, in Scenarios A to E.

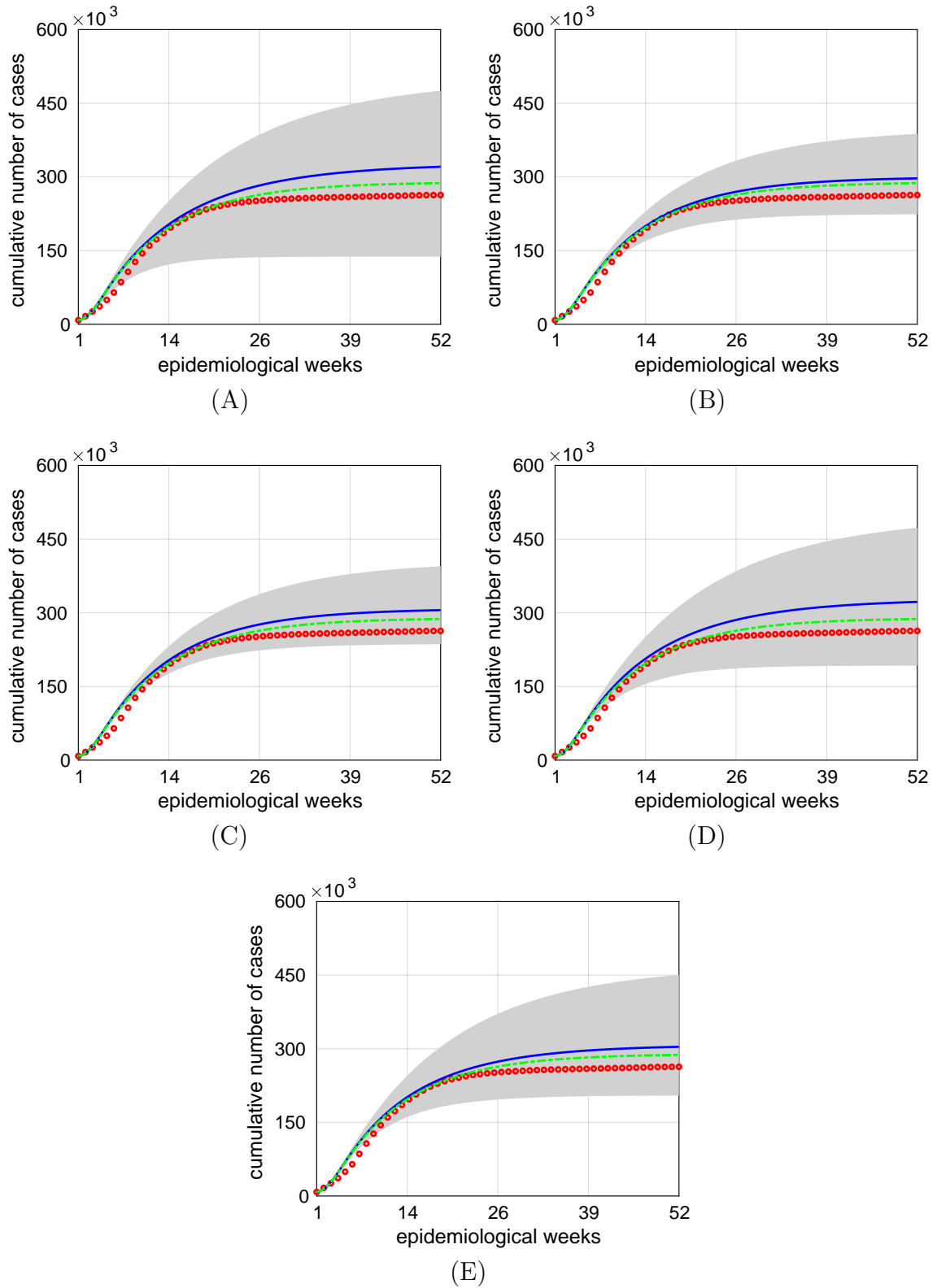


Figure 19 - The 95%-confidence bands for the model number of new cases, in Scenarios A to E.

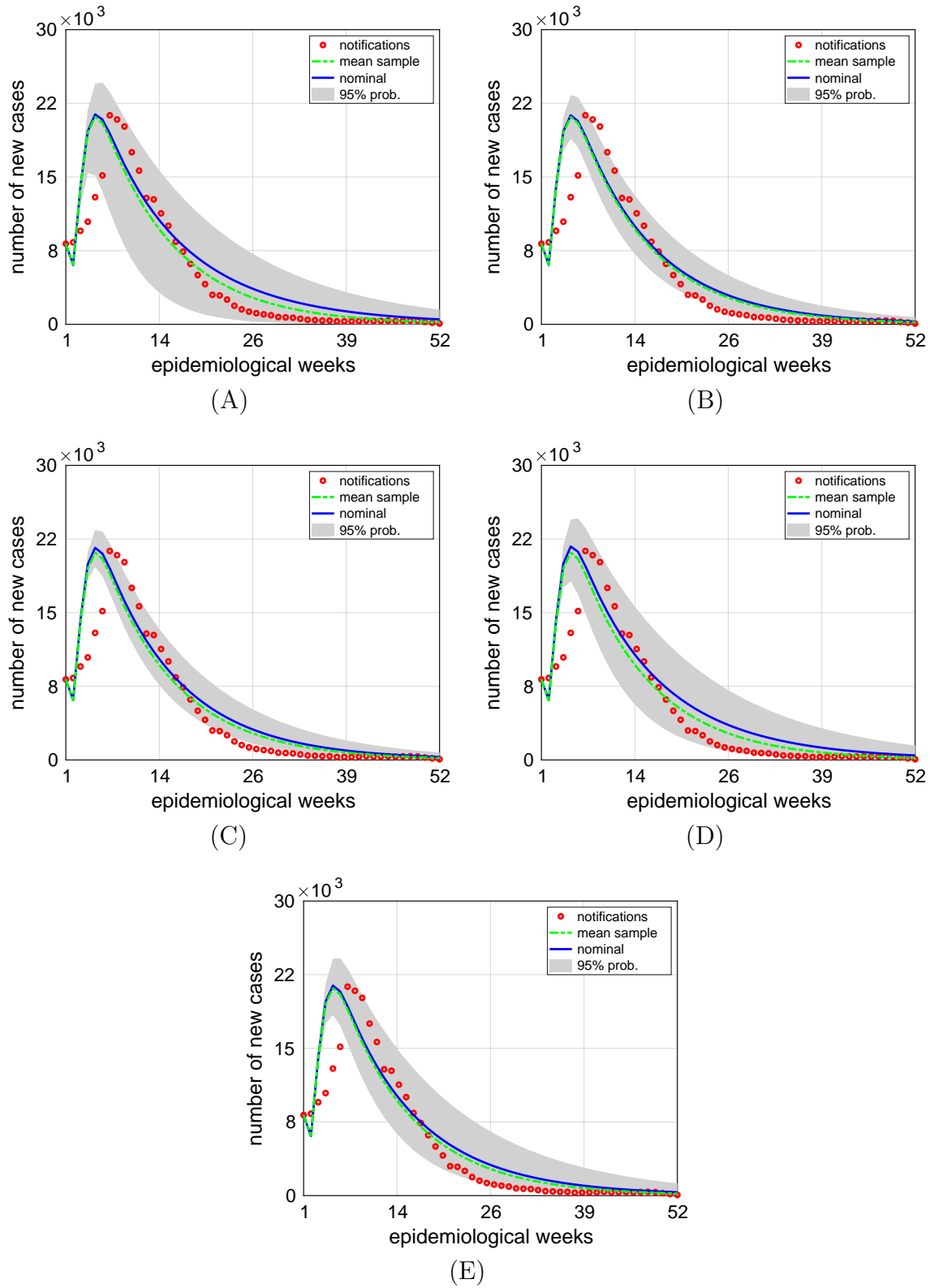


Figure 20 - Histogram and estimated PDF (kernel density) of the time average for the cumulative number of cases, in Scenarios A to E.

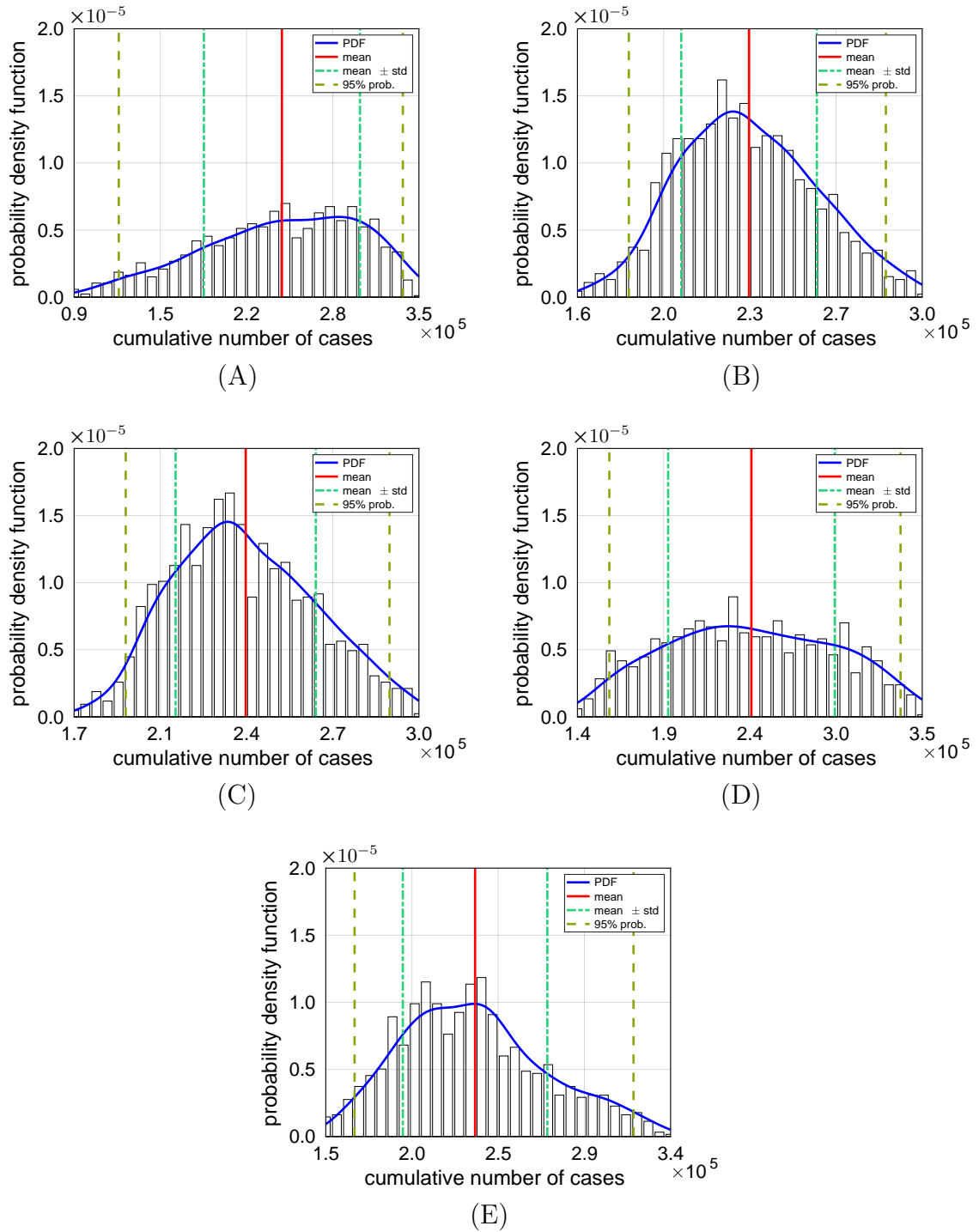


Figure 21 - Histogram and estimated PDF (kernel density) of the time average for the number of new cases, in Scenarios A to E.

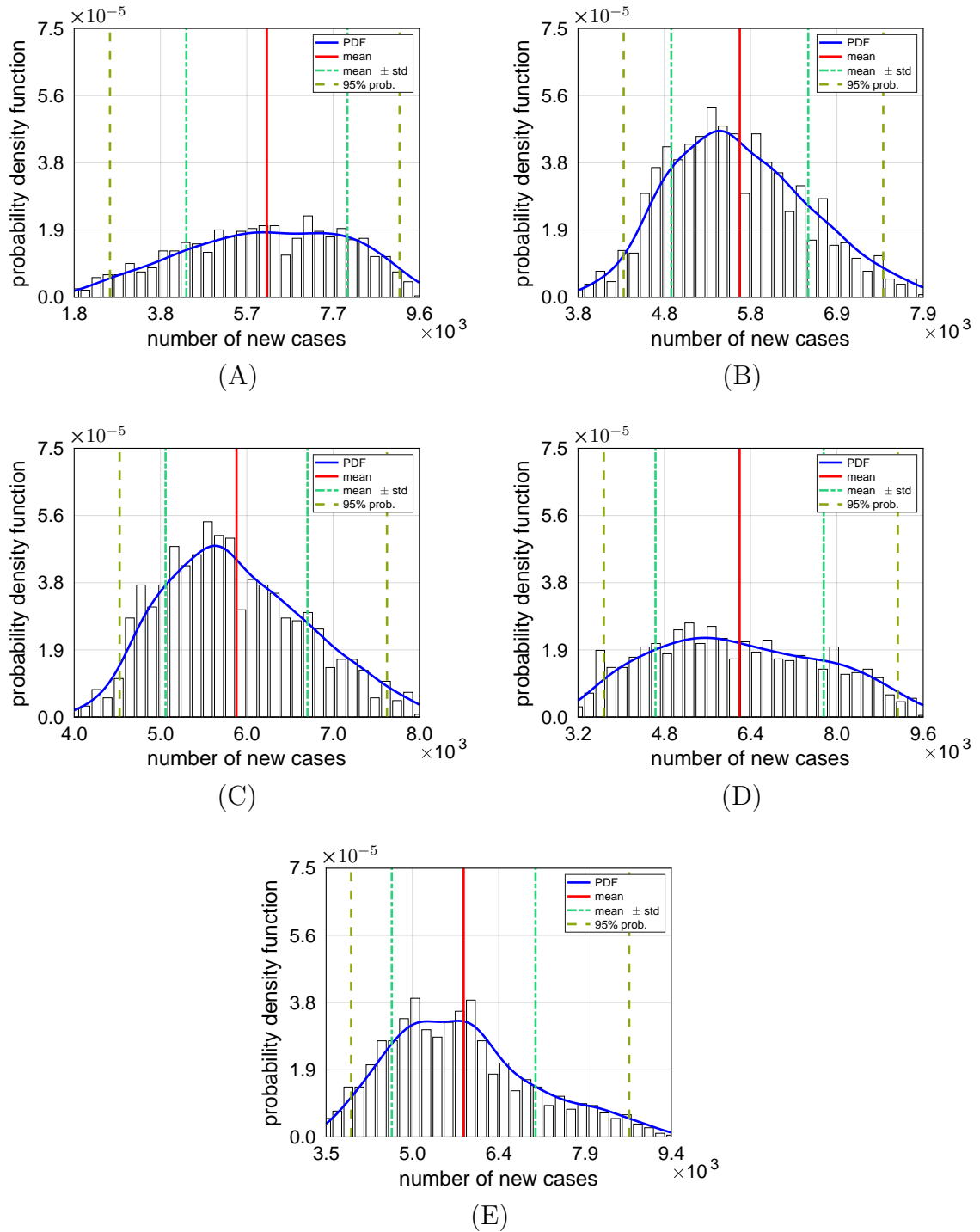
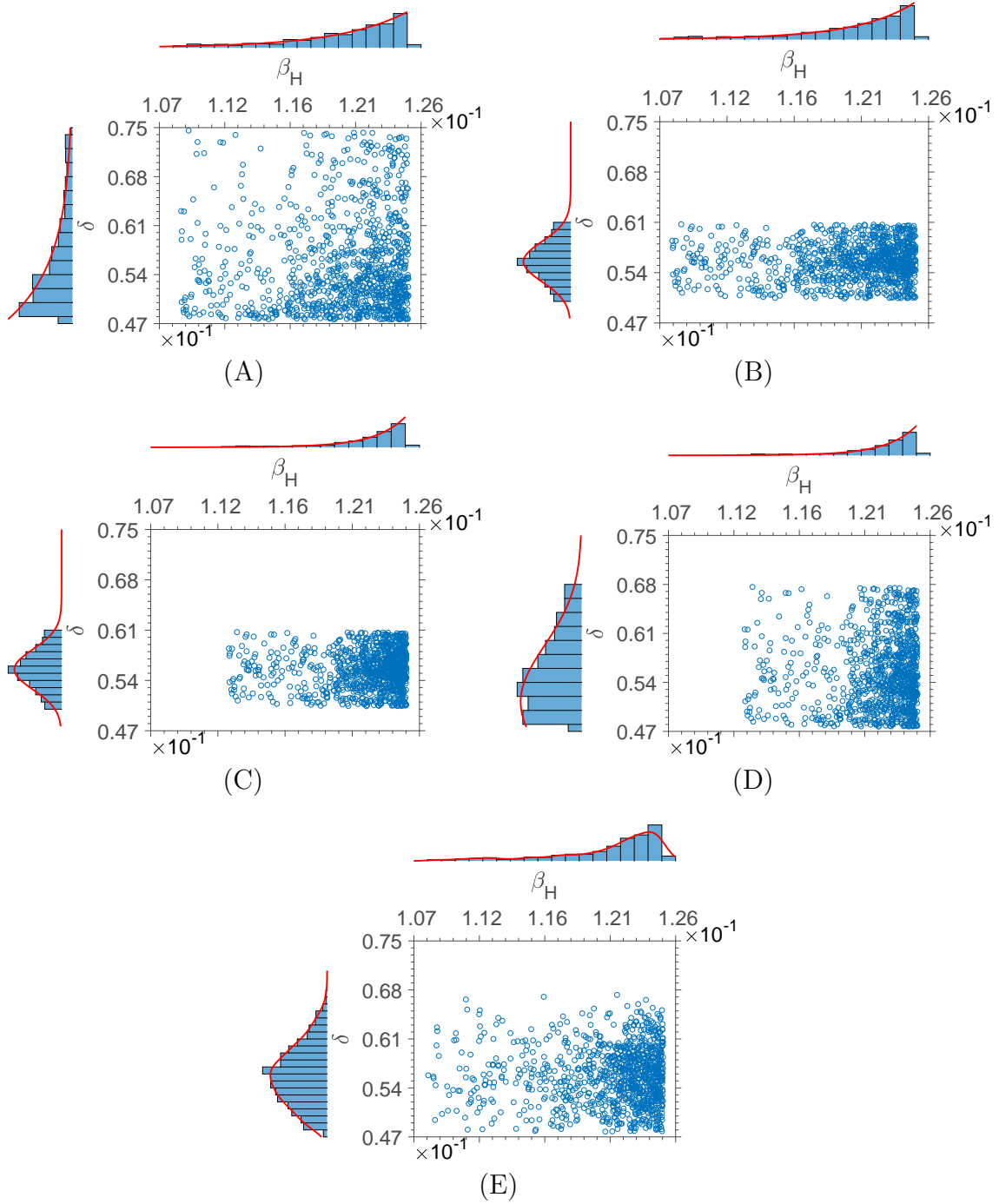


Figure 22 - Marginal distributions for the random inputs, in Scenarios A to E.



Additionally, another interest measure to be taken is the confidence interval for the attack rate (AR), important epidemiological quantity of reference. As the peak is where occurs the biggest contribution for the final size of the outbreak, the week (w_{\max}) and value of peak (\mathcal{N}_{\max}) will be putted on that analysis. All of these confidence intervals, for all the scenarios, are organized in the Table 4. Indeed, the analysis of the intervals support the argument that the effect of dispersion in δ was stronger. During the scenarios when CV_{δ} was smaller, the peak week was constant in 5, while it would be of greater interest for the interval to extend to EW 7. The AR intervals reveals that the outbreak occurred in Brazil was responsible to infect something between 0.1% and 0.2%, approximately. The proportion is very low, but the total population was very huge as well. Besides that, with a very small percentage of cases, the country are very susceptible to new outbreaks on the future. As the Brazilian authorities are now worry about the COVID-19 pandemic, probably a new Zika virus outbreak will occur very soon.

4.6 Some conclusions

In this chapter, the uncertainty quantification framework was applied to the 2016 Brazilian Zika virus outbreak using a double population compartmental model as tool to obtain the quantities of interest. The model response were calibrated to the Brazilian data in a previous work (DANTAS; TOSIN; CUNHA JR, 2018). So, the application of the framework started in the sensitivity analysis step. The Sobol indices reveal the pair of parameters β_h and δ as the most important on the first 20 EWs of the outbreak. The week of peak, instant of great interest, was discriminant on the analysis and main responsible for the pair selected of parameters. Through this information, a new model calibration could be conducted only adjusting the two most relevant inputs. The results obtained from this allowed to improve the response peak value, but not its position.

With the new calibration used as additional information about the parameters, multiple stochastic scenarios were tested, estimating the most unbiased parameter distributions with aid of the Maximum Entropy Principle for different values of dispersion. By that, the uncertainties were propagated to the system response in a Monte Carlo simulation. Confidence bands were calculated for the model QoIs and the histograms of time average were showed. To complement, confidence intervals for the attack rate, time of peak, and location of the peak were made too. The analysis reveal that increase on the input dispersion are more effective when applied to δ . Also, the proposed model performed robustness after the peak value, but reveal serious difficulties into capture the initial dynamics. An improvement can be search by including measures of model errors (MORRISON; CUNHA JR, 2020). So, this is the first immediate direction to follow in future works.

The estimated marginal distributions for β_h and δ in each scenario showed exponential distribution for δ only in the first scenario, where there is no informed dispersion. The parameter β_h manifested exponential distributions for four of the five studied scenarios, showing Gaussian shape in the last scenario alone.

Table 4 - 95% confidence intervals for the attack rate, peak value and peak location, in Scenarios A to E.

scenario	AR 95% CI	w_{\max} 95% CI	\mathcal{N}_{\max} 95% CI
A	$[0.067, 0.23] \times 10^{-2}$	[4, 6]	[15395, 24632]
B	$[0.11, 0.19] \times 10^{-2}$	[5, 5]	[18894, 23343]
C	$[0.11, 0.19] \times 10^{-2}$	[5, 5]	[19663, 23425]
D	$[0.093, 0.23] \times 10^{-2}$	[5, 6]	[18180, 24600]
E	$[0.099, 0.22] \times 10^{-2}$	[5, 6]	[18407, 24168]

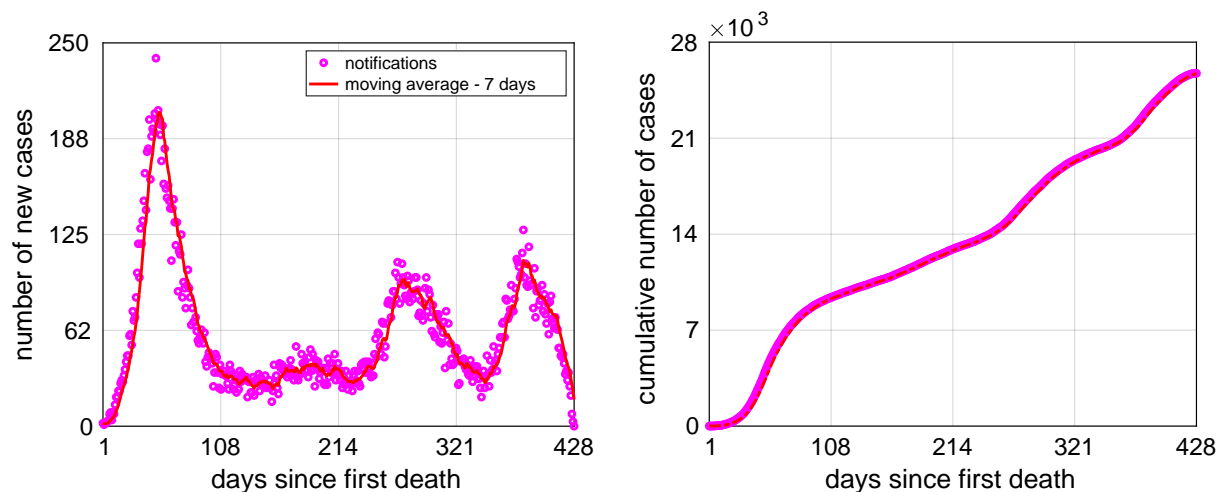
5 COVID-19 PANDEMIC IN RIO DE JANEIRO

The COVID-19 pandemic was as a public health emergency of international concern in January 2020 (WORLD HEALTH ORGANIZATION, 2020b). Even so, the first death detected in Rio de Janeiro (RJ) city was detected a couple months later, in March 17. Since then, the RJ outbreak had high and low moments, composing several waves. In this chapter, the RJ COVID-19 outbreak will be put under the microscope to analyze its comparative behavior depending on the wave observed. Again, the UQ framework will be applied to analyze how the proposed model performs when the inputs are subjected to uncertainties.

5.1 Data set

The data to be used in this chapter is the number of new deaths from COVID-19 in Rio de Janeiro city, since the first death detected (March 17, 2020) until May 20. The information was obtained by the Painel Rio COVID-19 (<https://experience.arcgis.com/experience/38efc69787a346959c931568bd9e2cc4>). The choice for using the deaths instead the cases is based on the asymptomatic behavior of the disease. Without a consolidated testing policy, that quantity itself is under several uncertainty. The deaths, on the other hand, are always detected. Of course, it will be always some errors, but probably much less than with the infections. The Figure 23 shows the referred data (and the cumulative form), with the moving average included to help observe the main evolution of the deaths.

Figure 23 - Time series for the daily number of new (left) and cumulative (right) deaths from COVID-19 in Rio de Janeiro since March 17, 2020.



5.2 Mathematical modeling

In this epidemiological scenario, different from the one presented in the Chapter 4, will be proposed a model uncoupled from the infection dynamics. It will be applied an multi waves Beta logistic growth model (BLM) to describe the cumulative number of deaths, D , following the equation

$$\frac{dD}{dt} = r D^q \left[1 - \left(\frac{D}{K} \right)^\alpha \right]^p, \quad (49)$$

where r is the growth rate, α is the asymmetry parameter, K is the carrying capacity, and q and p , respectively, controls the initial and final growths. The model QoI are the daily number of cumulative deaths, obtained by the integration the Eq. (49), and the daily number of new deaths

$$\mathcal{D} = D_d - D_{d-1}, \quad (50)$$

with d denoting the days, and $\mathcal{D}_1 = D_1$ for consistency. The multi-waves characteristic is inserted in the model by the consideration of time dependent parameters in the form

$$\zeta(t) = \zeta_1 + \sum_{i=1}^{n-1} \frac{\zeta_{i+1} - \zeta_i}{2} \left[1 + \tanh \left(\frac{\rho_i}{2} (t - \tau_i) \right) \right]. \quad (51)$$

The index n measure the number of waves defined, and the sub-parameters ρ and τ , respectively, represent the transition velocity and inflection point from each transition between waves. The model response will present as many waves as the parameters but the its waves is not necessarily synchronized. By the Figure 23, the time window is located in the interval $\{1, \dots, 428\}$, where the first and final days are March 17, 2020, and May 20, 2021, respectively. In that interval, the data shows a set of four waves. Because of this, a 4-waves model will be used. The computational model is assigned when are chosen the four individual values (one per wave) for $\{r_i, q_i, \alpha_i, p_i, K_i\}$, as well as the three values for the pair $\{\tau_j, \rho_j\}$, $j \in \{1, 2, 3\}$. Thus, the 4-waves BLM depends of 26 input sub-parameters. With these determined, the time evolution of set $\{r, q, \alpha, p, K\}$ can be calculated by the Eq. (51), allowing to obtain the daily number of cumulative and new deaths in the interval $[1, 428]$ through the Eqs. (49) and (50). As before, the integration step of this process is executed by the use of a Runge-Kutta (4,5) method. The initial condition used in all the simulations of this chapter is $C(1) = 2$, number of deaths reported in March 17.

5.3 Model calibration

Since there is no previous result of the proposed model for the COVID-19 outbreak in Rio de Janeiro, this section discusses a calibration process to obtain a more accurate model to be used in the next sections. The supports for the parameters were experimentally adapted from those discussed in (VASCONCELOS et al., 2021a). The carrying capacities are the only parameters that have the constraint of being monotonically increasing through the passing of waves.

Even with 428 data points to guide a tuning process, fit 26 parameters can be difficult. To simplify this, a sequential calibration is proposed to convert the original problem into a sequence of 4 calibration processes. The idea is to fit, in each case, the parameters associated with the wave of same number, while preserves the parameters obtained in previous solutions. With that, all the parameters are fitted recursively. In more details, the sequential calibration can be described by the following steps:

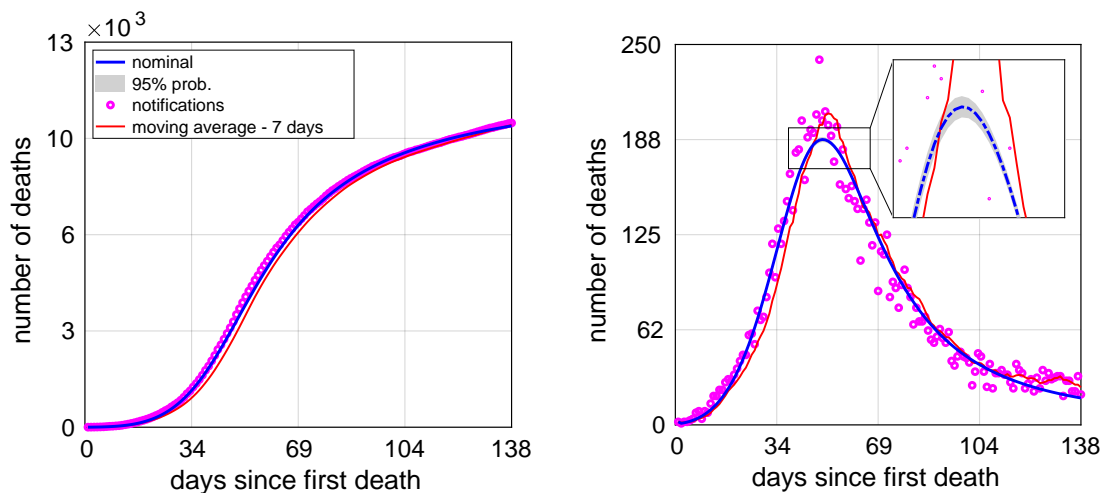
- (1) Calibrate the 5 parameters of an 1-wave BLM using the data from the first wave;
- (2) Use the parameters obtained in the $(n-1)$ previous calibrations to fit the remaining 7 parameters of a n -waves BLM, using the data from the n th first waves.

This recursive strategy is also useful for better defining the carrying capacities bounds. Since one of these are estimated, its value is used as lower bound for the next one. The upper bound is given for the cumulative value in the final of the wave from the data, increased in 2500 by the assumption of the wave have not reached its natural plateau before the transition for the next wave. This construction helps to maintain the consistency between each calibration step. Finally, the calibration process is performed using the cross-entropy method considering a Gaussian family distribution. The 1-norm is applied to measure the distance between the model daily number of new deaths response and the data, avoiding the effect of the outliers data points. The number of sample in each iteration was $N_s = 75$, and the elite percentile $\rho = 0.1$. A simple smooth is used in the update of the mean hyperparameter with $\zeta = 0.75$, and an iterative smoothing, characterized by $\theta = 0.9$ and $\vartheta = 7$, is applied during the iteration of the standard deviation. The stop criteria from each calibration process is $\text{tol}_\sigma = 0.01$ for the maximum standard deviation obtained in the actual iteration. As the parameters are from different orders of magnitude, it is included an additional stop criteria $\text{tol}_\mathcal{J} = 0.1$ for the improvement in the objective function ($|\mathcal{J}_i - \mathcal{J}_{i-1}|$) given by the actual iteration. Otherwise, the method stops when reaching 100 iterations. The data from the four waves is divided in: $\omega_1 = [1, 138]$, $\omega_2 = [139, 230]$, $\omega_3 = [231, 344]$ and $\omega_4 = [345, 428]$. The supports for the 26 parameters are listed in Table 5, together with the estimated values, means and standard deviation. All the calibrations ends by the objective function tolerance, after executing an average

of 60 iterations. The sequential results are showed in the Figures 24 – 27, where the 95% confidence bands are presented as well. The full calibration is displayed in the Figure 28.

The calibration figures demonstrate the efficiency of the sequential strategy. With less parameters to be fitted at the time, the convergence of the CE method occurs with no major complications. In the first 10 iterations, the method guide the response to a small dispersion ($CV \leq 5\%$) region already. The following iterations just update the values to the best fit under the tolerances described. Furthermore, it can be noted that the confidence bands for the cumulative numbers of deaths are larger for the last waves. This effect occurs because the error is propagated during more time, so the cumulative numbers get out of sync from the data points. The new deaths curves reveal that this does not compromise the quality of the waves fitting.

Figure 24 - First wave time series for the daily number of new (left) and cumulative (right) deaths in each wave from the first sequential calibration step.



5.4 Sensitivity analysis

As the model have the special quality of multi-waves behavior, this must be taken into account for sensitivity analysis proposes. In the previous chapter, the peak were considered as the most important instant of the outbreak. With several peaks, the reference will the waves itself. By the Eq. 51, each wave sub-parameters have an “on and off” effect during the evolution in time. However, its effect can be carried to the response even after being “turned off”. So, the cross-effect between parameters must be observed in each wave. In this section, the Sobol indices will be used to identified the most important parameters in each wave. For that, the number of new deaths will be calculated in 16

Table 5 - Lower and upper bounds for the 26 model parameters, and its respective estimation for the value, mean and standard deviation.

parameter	r_1	q_1	α_1	p_1	K_1		
lb	0	0	0	0	2		
ub	1	1	2	10	12771		
μ^*	0.5	0.78	1.46	2.49	12183		
σ^*	8×10^{-4}	3×10^{-4}	4.8×10^{-3}	4.9×10^{-3}	18		
\mathbf{x}^*	0.5	0.78	1.46	2.49	12183		
parameter	τ_1	ρ_1	r_2	q_2	α_2	p_2	K_2
lb	135	0	0	0	0	0	12183
ub	155	1	1	1	1	10	15786
μ^*	153	6.1×10^{-2}	0.5	0.65	0.68	0.78	15069
σ^*	1	1.2×10^{-3}	2.2×10^{-2}	4.8×10^{-3}	0.04	1.7×10^{-2}	102
\mathbf{x}^*	153	6.1×10^{-2}	0.46	0.65	0.69	0.76	15041
parameter	τ_2	ρ_2	r_3	q_3	α_3	p_3	K_3
lb	225	0	0	0	0	0	15041
ub	255	1	1	1	2	10	22643
μ^*	254	0.13	0.51	0.7	1.02	1.09	21973
σ^*	1	2.8×10^{-3}	0.02	4.1×10^{-3}	4.2×10^{-2}	1.8×10^{-2}	43
\mathbf{x}^*	255	0.13	0.51	0.7	0.97	1.08	21967
parameter	τ_3	ρ_3	r_4	q_4	α_4	p_4	K_4
lb	340	0	0	0	0	0	21967
ub	380	1	1	1	10	10	28205
μ^*	372	0.12	0.5	0.62	6.8	1.71	27236
σ^*	1	2.6×10^{-3}	3.4×10^{-2}	5.7×10^{-3}	0.2	0.07	58
\mathbf{x}^*	373	0.12	0.5	0.62	6.77	1.64	27231

Figure 25 - Second wave time series for the daily number of new (left) and cumulative (right) deaths in each wave from the second sequential calibration step.

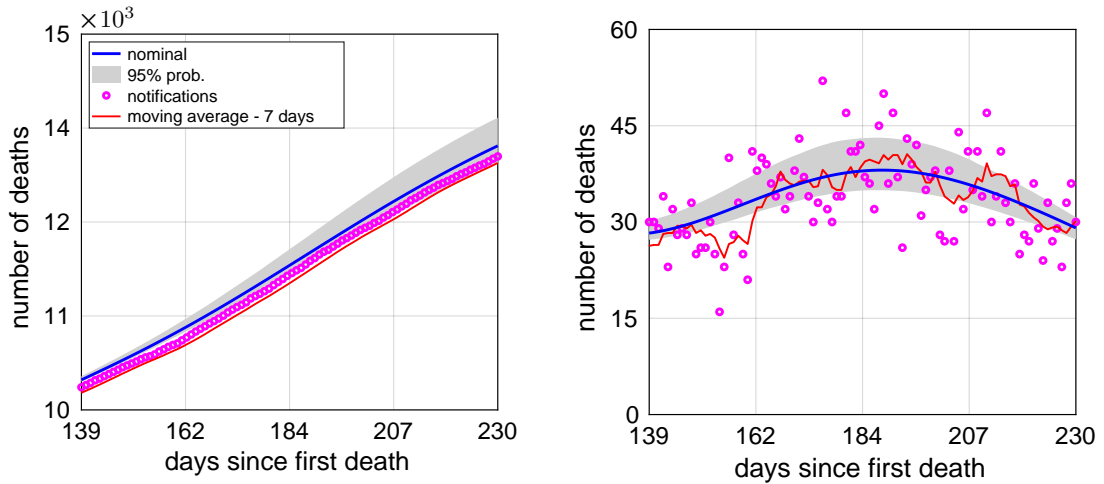
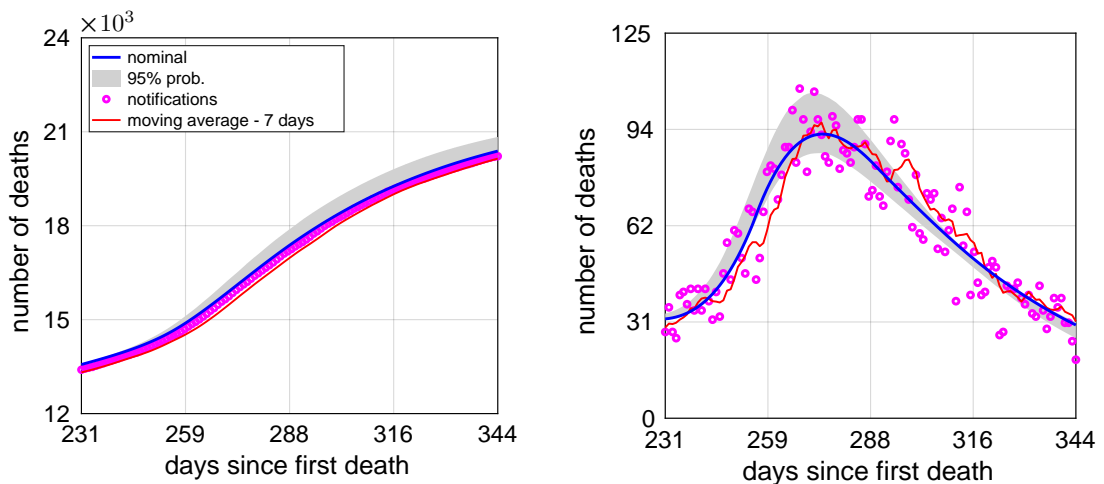


Figure 26 - Third wave time series for the daily number of new (left) and cumulative (right) deaths in each wave from the third sequential calibration step.



time instants equally spaced in each wave. The full 4-waves model will be used during all the analysis to not remove any possible cross-effect between the 26 parameters. Again, only the 10 most relevant set of indices will be displayed in each result.

First of all, the parameters distribution must be characterized. Since the tuned model proved to be very closer to the real data, the knowledge acquired during the calibration process can be used here. As described before, the first 10 iterations already guaranteed a coefficient of variation less or equal than 5%. So, using the final fitted parameters values as mean and a common $CV = 0.05$ to all of them, an uniform distribution can be created for each model parameters. That idea is to use the calibration process to

Figure 27 - Fourth wave time series for the daily number of new (left) and cumulative (right) deaths in each wave from the fourth sequential calibration step.

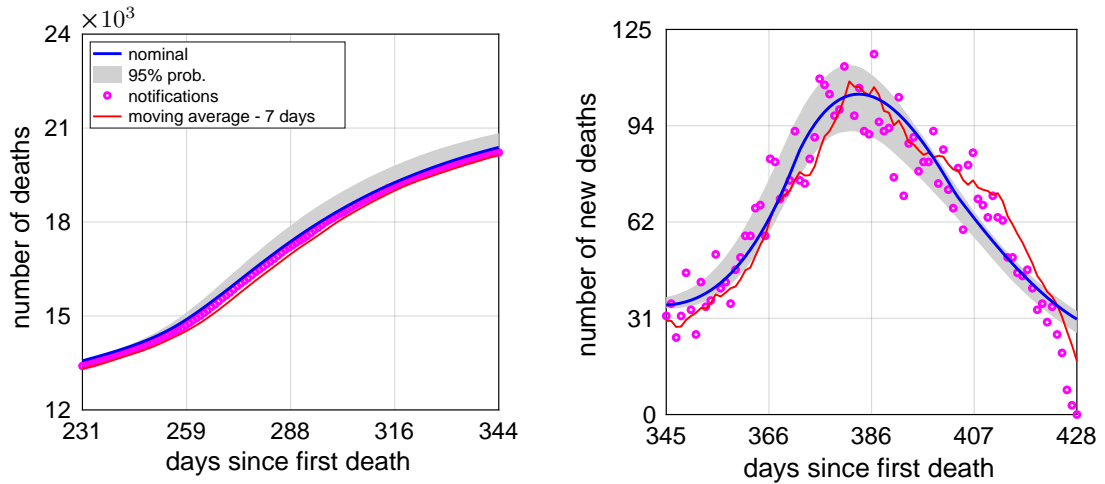
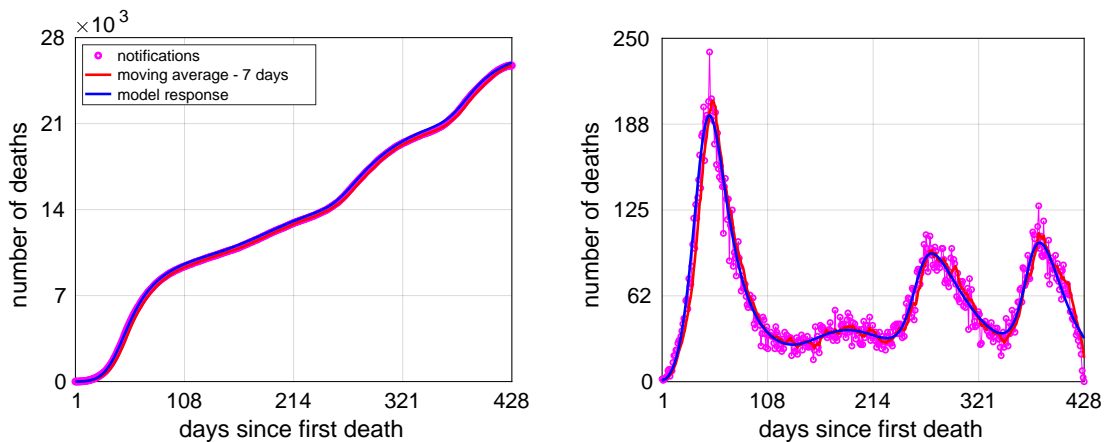


Figure 28 - Full time series for the daily number of new (left) and cumulative (right) deaths in each wave obtained through the sequential calibration process.



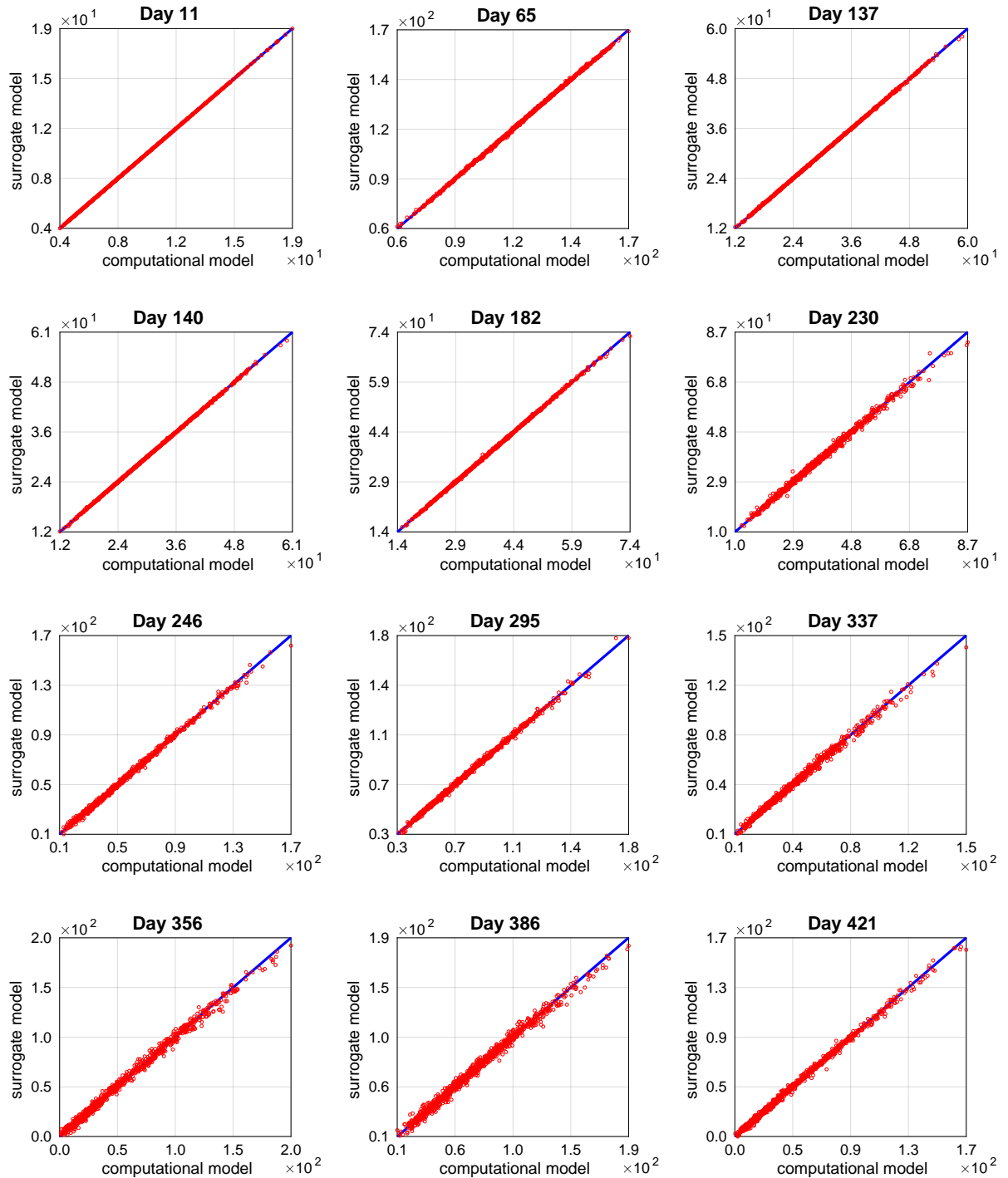
update the knowledge about the sub-parameters supports and reconstruct them. Therefore, all the model sub-parameters will be uniformly distributed on those new support.

With the random inputs characterized, the PCE will be constructed with a maximum polynomial degree of 5, in a total of 169911 coefficients to be obtained, and 2000 samples. Less samples are used here, compared to the Zika case, to maintain a good compromise between the accurate of the surrogate and computational cost, since the model QoI calculation depends of the integration of 5 hyperbolic tangents. To validate, a new set of 1000 samples was generated. The validation curves can be found in the Figures 29, where the most and the least accurate results in each wave are used to reference the general quality of the PCE created. The cross-validation error varied between the order of

10^{-8} for the better adjustments (first wave) and 10^{-2} for the worst cases (fourth wave). The other validation curves are reunited in the Appendix B.

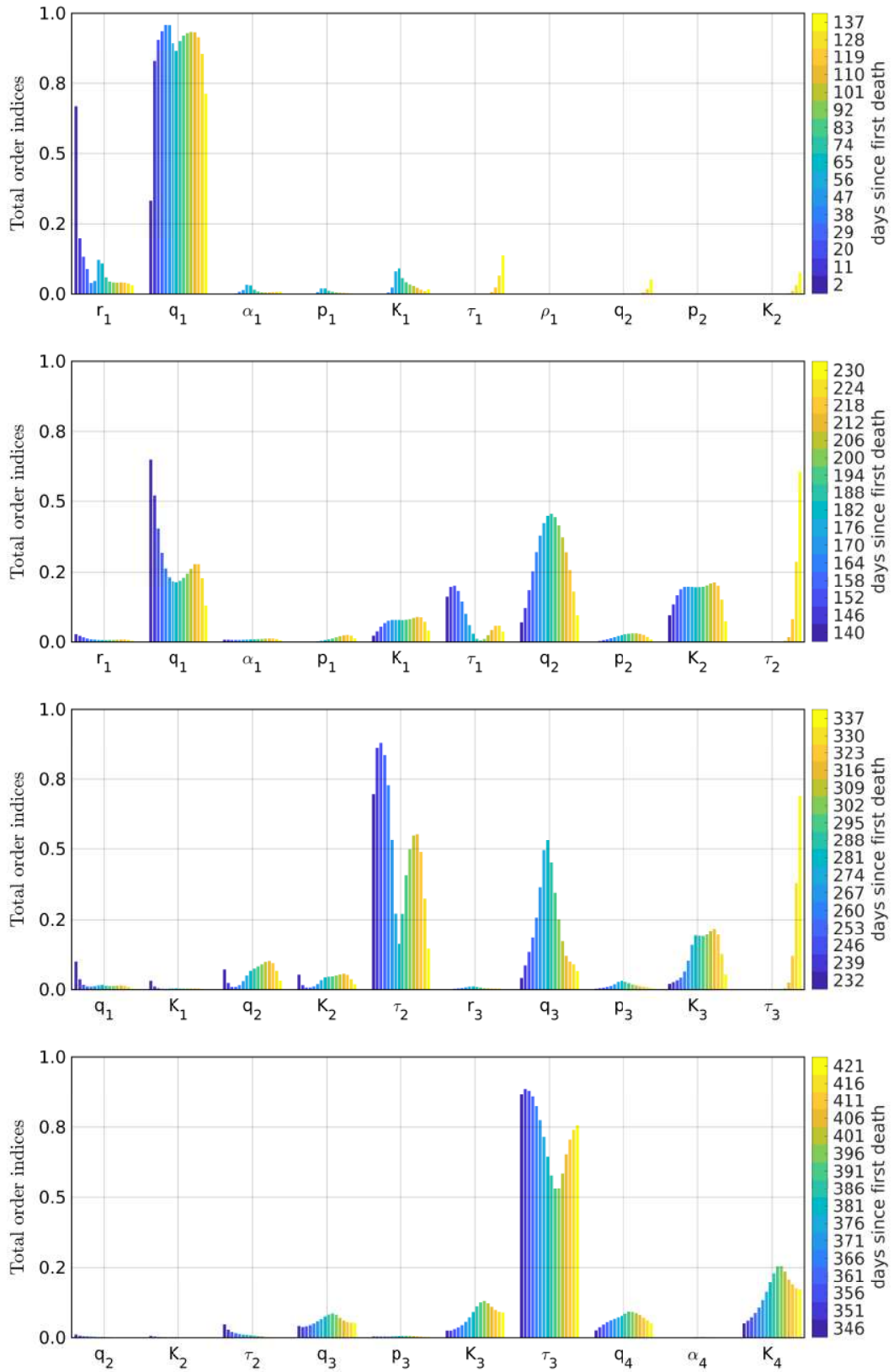
With the PCE model surrogate already created and validated, the Sobol indices can be calculated. Not all the possible orders will be necessary since the contributions strongly decrease in high orders. So, to focus only in the relevant orders while do not neglect possibly cross-effects in the total variance, the sensitivity indices will be calculated until the third order. Here, only the total indices figures will be exposed, with the others being reserved for the complementary results appendix. The intention is to compare parameters importance between waves. The omission of the individual orders does not generate loss for the present analysis, considering that the greatest contribution to the total indices comes from the first-order indices. The other orders slightly affect the quantitative point of view, but not the qualitative character of the analysis. The four total Sobol indices curves (one per wave) are put together in the Figure 30.

In the first wave, the initial growth control q_1 is protagonist. As the most of its indices is closer to 1, the parameter is obviously the most important in the wave. In second place, the parameters r_1 , K_1 and τ_1 are competitive with each other. Due to the initial effect, the growth rate r_1 will be selected as the second most relevant parameter. Then, from the analysis of the first wave, the pair $\{r_1, q_1\}$ are chosen as the most relevant. Even so, when including the second wave into the context, the parameters K_1 and τ_1 arise again. Although more discreet compared to q_1 , q_2 and K_2 , its effects can not be neglected. Pass to the triple $\{\tau_2, q_3, K_3\}$ mainly controls the variance. Finally, the fourth wave finish the analysis by indicating the parameters τ_3 and K_4 as the most relevant. In resume, the initial growth parameters (q_i), inflection times (τ_i) and carrying capacities (K_i) are the parameters that controls the variance in general. Depending of which i th wave is observed, i th (or the $(i - 1)$ th one) associated value for those three parameters will assume prominence. The exception is r_1 for the first wave. This occurs because in the first wave, the combined effect of the time variation in the model original parameters was not transferred to the response yet. So, as the growth rate, r_1 is very important at the moment. Also, by filtering only the 10 higher sets of indices to be analyzed, becomes clear how the contributions from the asymmetry parameters (α_i), final growth parameters (p_j) and the transition rates (ρ_i) for the model variance can easily be neglected with no great loss. In conclusion, by the analysis of the total Sobol indices in the four waves, from the 26 parameters, the most relevant are $\{r_1, q_1, K_1, \tau_1, q_2, K_2, \tau_2, q_3, K_3, \tau_3, K_4\}$. So, these 11 parameters must be taken into consideration for the uncertainty propagation studies. Seems a lot to keep so much parameters, when compared with the Zika study where only two parameters were selected from the original set of six (one-third reduction). However, here the multi-wave characteristic of the present scenario must be considered. If reducing the set of 11 parameters, relevant effects will be loss of some of the waves. Now is preferable to keep more parameters so as not to benefit one wave over another.

Figure 29 - Validation plots for the PCE surrogate constructed for the QoI \mathcal{D} .

Legend: The red dots represent the sample and the blue line is the identity line.

Figure 30 - Total Sobol indices for the daily number of new deaths per wave.

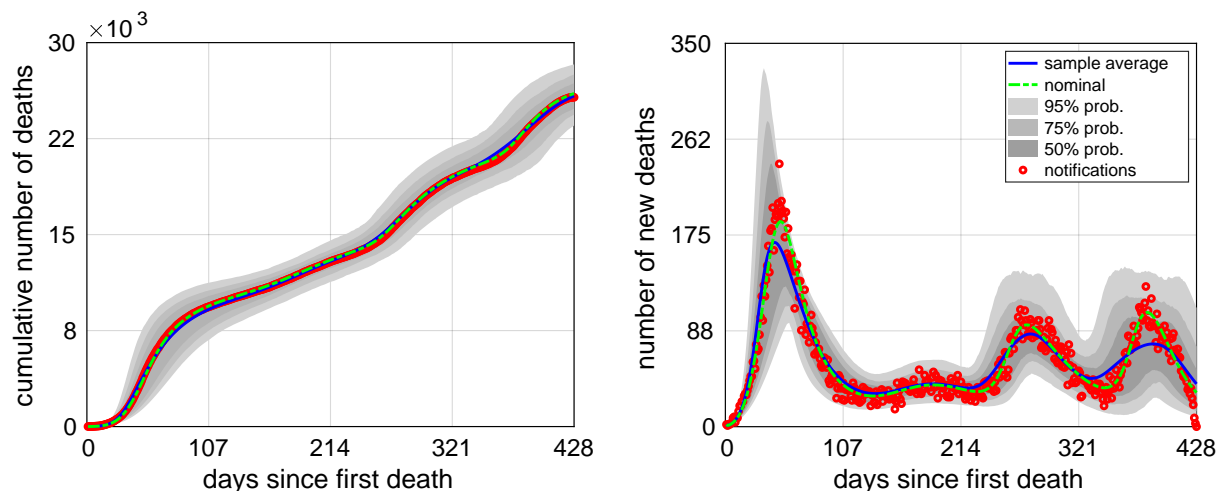


5.5 Uncertainty propagation

After the sensitivity analysis step, now the uncertainties from the 11 parameters selected can be executed. As before, will assume that the calibration allows to update the parameters bounds by considering a coefficient of variation of 5%. Thus, the same uniform distribution construction will be made. But here, the parameters which were not selected in the Sobol indices analysis will be considered as constant, with value equal to the found in the calibration (Table 5). The idea is to analyze the model robustness by studying its capacity of describe not only the average trend of the data, but also the data points itself. Of course, the real data presents a great dispersion. If the model is able to capture this dispersion, with small variability in the inputs, it is added a new degree of validation for its use in complementary studies on the phenomena.

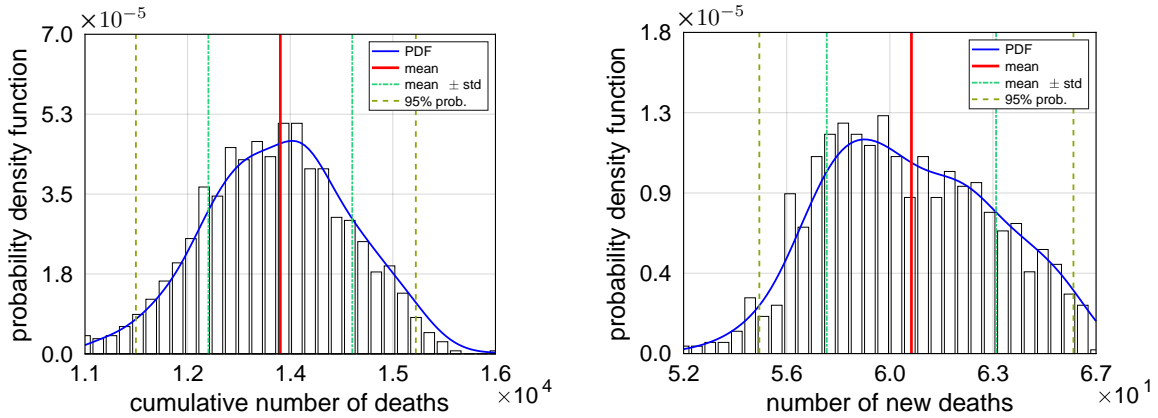
The Figure 31 shows the 95%, 75% and 50% confidence bands for the model QoIs. The 95% band can encapsulate almost all the data points, with exception of some points that clearly look like outliers. When considering the 75% bands, the most part of the data is cover, showing how the model is effective into it. The number of points marked by the 50% is obviously smaller than for other envelopes, but, still, the points that seems to guide the trend are covered. In conclusion, when considering the uncertainties, the model well describe all the waves of the studied scenario. To complement the uncertainty quantification study, the time average histograms for the model cumulative number of deaths and number of new deaths are showed in the Figure 32.

Figure 31 - The 95%-confidence bands for the model cumulative number of cases (left) and number of new deaths (right).



As the 4-wave BLM used in this chapter part of the premise of considering the original 5 parameters as time variants, its evolution with the addition of the uncertainties

Figure 32 - Histogram and estimated PDF (kernel density) of the time average for the model cumulative number of cases (left) and number of new deaths (right).



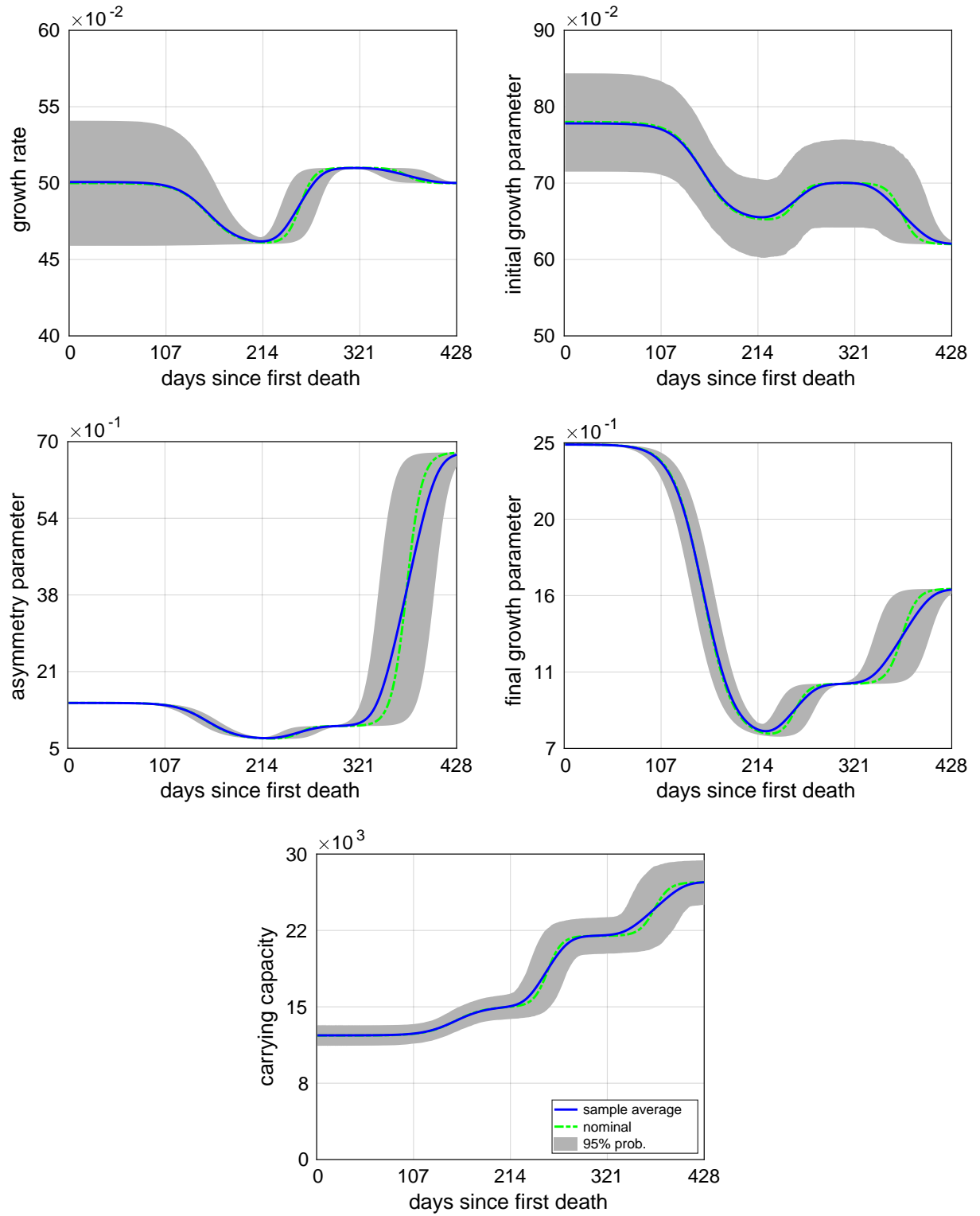
can be observed. Of course, for the parameters there is no data to be compare. So, it is not necessary to create three different confident bands for each one, as was done for the QoIs. The 95% confident bands generated for the model parameters are reunited in the Figure 33. This result is interesting because the connection from which parameters were considered random is easy to be made. The random variables q_1 and K_1 , for example, will not affect the evolution of $r(t)$. Moreover, as all the carrying capacity parameters (K_1, K_2, K_3, K_4) are random, the confidence band from $K(t)$ is, clearly, the only one that does not have the taper regions. There occurs in the moments where a constant parameters is controlling the time evolution. For example, the taper regions in $r(t)$ starts after r_1 take place to r_2 . The envelope does not fully agglutinate because of the presence of the τ_1 and τ_2 which are both random. An analogous situation can be found for $p(t)$ and $\alpha(t)$.

To finish the discussion, other useful information to be extracted from the Monte Carlo uncertainty propagation simulation are estimates from the number of deaths due each wave, that is, that gain in the cumulative number of deaths from a wave to the next one. Using the 95% envelope in the final time instants of each wave, the Table 6 reunites the estimated confident intervals for the increase of deaths.

Table 6 - Estimated 95% confident intervals for the final cumulative number of deaths and total increase of deaths, for each wave.

wave	1	2	3	4
95% CI total of deaths	[8332,11813]	[11871,15141]	[18200,22814]	[23520,28300]
95% CI gain of deaths	[8332,11813]	[58,6809]	[3059,10943]	[706,10100]

Figure 33 - The 95%-confidence bands for the model parameters.



5.6 Some conclusions

In this chapter was presented a 4-waves Beta logistic model to describe the 2020-2021 (until May 20) COVID-19 pandemic in Rio de Janeiro city, Brazil. The multi waves behavior is injected in the model response by considering the 5 original parameters as time dependents, following a smooth formulation based in hyperbolic tangents. With that, the model have a total of 26 input sub-parameters.

The uncertainty quantification framework were applied to the proposed model. First of all, to fit the model inputs, a sequential calibration design exploring cross-entropy method is proposed. It splits the original tuning problem in a set of 4 smaller problems focused into fit specific sets of parameters, using the parameters already calibrated as constant. The results show that the sequential strategy is very effective to fit this relatively high number of model inputs.

The iteration of the cross-entropy method on the proposed model helps to construct region of 5% dispersion around the fitted model sub-parameters. Following the framework steps, the global sensitivity analysis was performed on that region to inform which inputs are more relevant. The importance of each one changes depending on the wave observed. By comparing the four waves, 11 inputs were selected from the original 26. In general the parameters of final growth and carrying capacity, domains the model variance, together with the transition time parameters. An uncertainty propagation simulation was made to see how the selected sub-parameters affects the model outputs. Through the 95%, 75% and 50%, confidence bands it was possible to visualize how the model capture the data dispersion. With lower dispersion on the inputs, the 50% band cover the main density of data. The two other bands capture the most part of the outlier data points. The effect of the random inputs could be observed more isolated on the confidence bands constructed for the 5 time varying model parameters.

The 95% confidence bands for the model QoI and time evolution of the parameters were show. Also, time average histogram for the model outputs were displayed. The uncertainty propagation study reveals the robustness of the proposed model into describe the real data. By the analysis of the parameters bands, the effect given for each random input could be observed more isolated. In addition, intervals for the increase of deaths due to each wave were estimated. By that, the first wave is naturally the one whose more deaths occurs. The third wave is the second higher on that referenced measure.

6 FINAL REMARKS

6.1 Research contributions

The project covers the study of epidemiological phenomena, its mathematical modeling and analysis on the presence of uncertainties. To that, an UQ framework was presented. It combines PCE-based Sobol indices global sensitivity analysis to identify the most important input parameters, with Maximum Entropy Principle to construct unbiased distributions. So, Monte Carlo simulations allows to analyze the statistics of the outcome stochastic process.

The framework was applied in two different recent outbreak scenarios: The 2016 Zika virus outbreak in Brazil and the the 2020-2021 COVID-19 outbreak in Rio de Janeiro city. For the first, the model used was a double population compartmental model. In the second one, a 4-waves Bela logistic model (BLM) was explored.

In the Zika scenario, sensitivity analysis results allowed to select the two most important inputs, reducing the amount of random variables to be explored in the uncertainty propagation to one third. The Sobol indices were also used to improve the previous calibration result to a new one that betters capture the peak value. Five scenarios of uncertainties were develop to study how the model performs on different values of known input dispersion. Complementary statistics for the model QoIs, time of peak, value of peak and attack rate, were demonstrated.

For COVID-19 in Rio de Janeiro, a sequential cross-entropy calibration process was design to tuning the BLM to the real data, by dividing the original problem in a set of 4 minor problems focuses in specific sub sets of the parameters, that recursively uses the previous solutions. The calibration result was explored to update the parameters supports. A sensitivity analysis study on these new supports was conducted to identify the most relevant model inputs in different wave. After, an uncertainty propagation simulation on the 11 important parameters was made. Three different confident bands (95%,75%,50%) were constructed to help observe the model capacity into cover the data. Other statistics were exposed and intervals for the gain of deaths due each COVID-19 wave could be estimated.

6.2 Main conclusions

The analysis from the Zika scenario reveal the β_h and δ parameters as the most important in the first 20 EW of the outbreak. The improved calibration corrected the peak value but was not capable of adjust its position, still occurring before the real

one. The uncertainty propagation scenarios studied reveal that higher dispersion is more effective to the response variability when applied to the parameter δ than to β_h . Also, the study allowed conclude that the compartmental model is robustness to reproduce the Brazilian outbreak on presence of uncertainties only after the peak (EW 7), but have serious difficulties into capture the initial dynamics.

About the COVID-19 scenario, the 4w-BLM calibrated by the sequential strategy was very closer to the original the real data from the Rio de Janeiro deaths waves. Through the Sobol indices analysis, the final growth sub-parameters (q) showed up consistently high important during all the waves. The carrying capacities and transition times are also relevant and its effect increase during the passing of waves. By simulating uncertainty propagation scenario on the main 11 sub-parameters it was possible to infer that the model is robust to describe the real deaths. With small dispersion on the inputs the model response capture the most part of the data points inside the 50% confidence band. The 75% and 95% bands cover almost all the outliers. The confidence intervals for the death gains reveal the first and third waves as the two which more contributed to the cumulative amount of lost lifes.

6.3 Future directions

A first direction will be to adapt the uncertainty quantification framework to guide a study of model errors, where the sensitivity analysis will be used to guide which kind of error are more effective to make up for the deficits found here. The idea is intended to be applied to continues the studies on the Zika scenario.

For the scenario of COVID-19, the next step is to explore the calibrated model to test some case studies during the waves and how some control measures could be used to contain it. Also, an idea to be explored is to apply the model for the full Brazilian deaths. Then, its robustness can be put on prove on the light of the national scale scenario, way less isolated.

To finish, it is desired to develop the ideas of how to use sensitivity analysis to improve previous resolves. The Sobol indices model selection criterion presented in the Appendix A should be applied on other studies where the results could be compared the those obtained for other established methods to verify the advances on use the sensitivity analysis-based methodology.

6.4 Scientific production and events

During Master's degree period, some scientific works were developed and presented in events of wide interest. In 2019, some of the results on the Zika scenario were presented in the Conference of Computational Interdisciplinary Sciences (CCIS 2019). The work was later published in the Journal of Computational Interdisciplinary Sciences (DANTAS; TOSIN; Cunha Jr, 2019). The results not covered in this paper are reunited in a more complete work, currently under review for submission in a research journal. Poster contributions around it were took to the Conference on Perspectives in Nonlinear Dynamics (PNLD 2019). At the end of the year, the master student made a presentation about construction of models and model errors in the *XIV Conferência Brasileira de Dinâmica, Controle e Aplicações* (XIV Brazilian Conference of Dynamics, Control and Applications, in english) (DINCON 2019).

In 2020, the quarantine from COVID-19 starts and affects the academic events. Some of them were canceled, postponed or moved to online mode. So, contributions were presented only in the *Congresso Brasileiro de Automática* (Brazilian Congress of Automatics, in english) (CBA 2020). There, was presented the preliminary results on Sobol indices-based model selection. Also, before the quarantine starts, a book chapter describing the calculation of the PCE-based Sobol indices for biological phenomena was accepted for publication (TOSIN; CÔRTES; CUNHA JR, 2020).

With the pandemics, the search for packages easy-to-run to simulate the dynamics of infectious diseases increases. To help in this sense, there was developed the EPIDEMIC - Epidemiology Educational Code - , an educational Matlab toolkit for epidemiological analysis (PAVLACK et al., 2021). Also, the routines used to analyze the Zika scenario were organized in the ZIKAVD – Zika Virus Dynamics – package, to facilitate the reproduction and adaptation for other similar outbreak (TOSIN; DANTAS; CUNHA JR., 2021).

On the COVID-19 results, a paper about it has been develop and a work was already accepted to presentation in the 3rd Pan American Congress on Computational Mechanics (III PANACM), in the end of 2021.

To complement his theoretical background on dynamic systems and modeling of biological process, the dissertation's author also participate in several events as listener. In particular, he recently was remotely present in the Encontro nacional de modelagem matemática da Covid-19 (National encounter on mathematical modeling of the Covid-19) (ENMM-Covid19). Before the quarantine starts, the student also were selected to participates in two courses occurred in the city of São Paulo: The São Paulo School of Advanced Sciences on Nonlinear Dynamics (SPNL 2019) and the IX Summer School on Mathematical Biology (SSSMB 2020).

REFERENCES

- AHRENS, W.; PIGEOT, I. (Ed.). *Mathematical epidemiology*. 2nd. ed. New York: Springer, 2014.
- AKAIKE, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, v. 19, n. 6, p. 716–723, 1974.
- ALSAFI, M. Nicolaa abd Z. et al. The socio-economic implications of the coronavirus pandemic (COVID-19): A review. *International Journal of Surgery*, v. 78, p. 185–193, 2020.
- ARONNA, M. S.; GUGLIELMI, R.; MOSCHEN, L.M. A model for COVID-19 with isolation, quarantine and testing as control measures. *Epidemics*, v. 34, p. 100437, 2018.
- ATALAN, A. Is the lockdown important to prevent the COVID-19 pandemic? Effects on psychology, environment and economy-perspective. *Annals of Medicine and Surgery*, v. 56, p. 38–42, 2020.
- A.TSOULARIS; J.WALLACE. Analysis of logistic growth models. *Mathematical Biosciences*, v. 179, n. 1, p. 21–55, 2002.
- BACAER, N. Verhulst and the logistic equation (1838). In: *A short history of mathematical population dynamics*. London: Springer, 2011. p. 35–39.
- BARRETO, M. L.; TEIXEIRA, M. G.; CARMO, E. H. Infectious diseases epidemiology. *Journal of Epidemiology and Community Health*, v. 60, n. 3, p. 192–195, 2006.
- BASSANEZI, R. C. *Temas e modelos*. 1^a. ed. Campinas: Editora Unicamp, 2012.
- BEGON, M. et al. A clarification of transmission terms in host-microparasite models: numbers, densities and areas. *Epidemiology and Infection*, v. 129, n. 1, p. 147–153, 2002.
- BLACK, S. et al. Transforming vaccine development. *Seminars in Immunology*, v. 50, p. 101413, 2020.
- BLATMAN, G.; SUDRET, B. An adaptive algorithm to build up sparse polynomial chaos expansions for stochastic finite element analysis. *Probabilistic Engineering Mechanics*, v. 25, n. 2, p. 183–197, 2010.
- BLOOM, D. E.; BLACK, S.; RAPPUOLI, R. Emerging infectious diseases: A proactive approach. *Proceedings of the National Academy of Sciences of the United States of America*, v. 114, n. 16, p. 4055–4059, 2002.
- BLUMBERG, A. A. Logistic growth rate functions. *Journal of Theoretical Biology*, v. 21, n. 1, p. 42–44, 1968.
- BOTEV, Z. I. et al. The cross-entropy method for optimization. In: *Handbook of Statistics 31*. Oxford: Elsevier, 2007. v. 31, p. 35–59.
- BRAUER, F. Mathematical epidemiology: Past, present, and future. *Infectious Disease Modeling*, v. 2, n. 2, p. 113–127, 2017.

- BRAUER, F.; CASTILLO-CHAVEZ, C.; FENG, Z. (Ed.). *Mathematical models in epidemiology*. New York: Springer, 2019.
- BRAUER, F.; VAN DEN DRIESSCHE, P.; WU, J. (Ed.). *Mathematical epidemiology*. Berlin: Springer, 2008.
- CHEN, D.-G.; CHEN, X.; CHEN, J. Reconstructing and forecasting the COVID-19 epidemic in the United States using a 5-parameter logistic growth model. *Seminars in Immunology*, v. 5, n. 25, 2020.
- COSTA, A.; JONES, O. D.; KROESE, D. Convergence properties of the cross-entropy method for discrete optimization. *Operations Research Letters*, v. 35, n. 5, 2007.
- COSTA, G.; COTA, W.; FERREIRA, S. C. Outbreak diversity in epidemic waves propagating through distinct geographical scales. *Physical Review Research*, v. 2, n. 4, p. 043306, 2020.
- CUNHA JR, A. Modeling and quantification of physical systems uncertainties in a probabilistic framework. In: *Probabilistic Prognostics and Health Management of Energy Systems*. Cham: Springer, 2017. p. 127–156.
- _____. Enhancing the performance of a bistable energy harvesting device via the cross-entropy method. *Nonlinear Dynamics*, v. 103, p. 137–155, 2021.
- CUNHA JR, A. et al. *Relatorio 02, progresso da COVID-19 no Brasil e no estado do Rio de Janeiro: 22ª semana epidemiológica do calendário 2020 (24/5/2020 até 30/5/2020)*. 2020.
- CUNHA JR, A. et al. Uncertainty quantification through Monte Carlo method in a cloud computing setting. *Computer Physics Communications*, v. 185, n. 5, p. 1355–1363, 2014.
- DANIEL, S. J. Vaccine hesitancy, vaccine refusal and the anti-vaccine movement: influence, impact and implications. *Prospects*, v. 49, p. 91–96, 2020.
- DANTAS, E.; TOSIN, M.; CUNHA JR, A. Calibration of a SEIR–SEI epidemic model to describe the Zika virus outbreak in Brazil. *Applied Mathematics and Computation*, v. 338, p. 249–259, 2018.
- DANTAS, E.; TOSIN, M.; Cunha Jr, A. An uncertainty quantification framework for a Zika virus epidemic model. *Journal of Computational Interdisciplinary Sciences*, v. 10, n. 91, 2019.
- DANTAS, E. et al. A mathematical analysis about Zika virus outbreak in Rio de Janeiro. *Journal of Computational Interdisciplinary Sciences*, v. 9, n. 1, p. 249–259, 135.
- DE GROOT, R. J. et al. Middle East respiratory syndrome coronavirus (MERS-CoV): Announcement of the coronavirus study group. *Journal of Virology*, v. 87, n. 14, p. 7790–7792, 2013.
- DE JESUS, J. G. et al. *First cases of coronavirus disease (COVID-19) in Brazil, South America (2 genomes, 3rd March 2020)*. 2020. (<https://virological.org/t/first-cases-of-coronavirus-disease-covid-19-in-brazil-south-america-2-genomes-3rd-march-2020/409>).

- DOHERTY, P. C. *Pandemics: What everyone needs to know*. Oxford: Oxford University Press, 2013.
- DONALISIO, M. R.; FREITAS, A. R. R.; von Zuben, A. P. B. Arboviruses emerging in Brazil: challenges for clinic and implications for public health. *Revista de Saúde Pública*, v. 51, n. 30, p. 1–6, 2016.
- DOS SANTOS, T. et al. Zika virus and the Guillain-Barré syndrome – Case series from seven countries. *New England Journal of Medicine*, v. 375, n. 16, p. 598–1601, 2016.
- DUBÉ, E.; VIVION, M.; MACDONALD, N. E. Vaccine hesitancy, vaccine refusal and the anti-vaccine movement: influence, impact and implications. *Expert Review of Vaccines*, v. 14, n. 1, p. 99–117, 2015.
- GENTLE, J. E.; HÄRDLE, W. K.; MORI, Y. (Ed.). *Handbook of Computational Statistics*. 2nd. ed. Berlin: Springer, 2012.
- GHANEM, R.; Red-Horse, J. Polynomial chaos: Modeling, estimation, and approximation. In: *Handbook of Uncertainty Quantification*. Cham: Springer, 2017. p. 521–551.
- GHANEN, R. G.; SPANOS, P. D. (Ed.). *Stochastic Finite Elements: A Spectral Approach*. Revised. New York: Dover publications, 2012.
- GIESECKE, J. *Modern Infectious Disease Epidemiology*. 3rd. ed. Florida: CRC Press, 2017.
- GLATTER, K. A.; FINKELMAN, P. History of the plague: An ancient pandemic for the age of COVID-19. *The American Journal of Medicine*, 2020.
- GRIMMETT, G.; WELSH, D. *Probability: An Introduction*. 2nd. ed. Oxford: Oxford University Press, 2017.
- GUAN, Y. et al. The emergence of pandemic influenza viruses. *Protein and Cell*, v. 1, n. 1, p. 9–13, 2010.
- GUARNER, J. Three emerging coronaviruses in two decades. *American Journal of Clinical Pathology*, v. 153, n. 4, p. 420–421, 2020.
- HAYS, J. N. *Epidemics and pandemics: Their impacts on human history*. California: ABC-CLIO, 2005.
- HOULIHAN, C. F.; WHITWORTH, J. A. G. Outbreak science: recent progress in the detection and response to outbreaks of infectious diseases. *Clinical Medicine*, v. 19, n. 2, p. 140–144, 2019.
- HUANG, Y.-J. S.; HIGGS, S.; VANLANDINGHAM, D. L. Emergence and re-emergence of mosquito-borne arboviruses. *Current Opinion in Virology*, v. 34, p. 104–109, 2019.
- JAYNES, E. T. Information Theory and Statistical Mechanics. *Physical Review*, v. 106, n. 4, p. 620–630, 1957.
- KAIN, T.; FOWLER, R. Preparing intensive care for the next pandemic influenza. *Critical Care*, v. 23, n. 337, p. 1–9, 2019.

- KEELING, M. J.; EAMES, K. T. D. Networks and epidemic models. *Journal of the Royal Society Interface*, v. 2, n. 4, p. 295–307, 2005.
- KEELING, M. J.; ROHANI, P. *Modeling Infectious Diseases in humans and animals*. New Jersey: Princeton University Press, 2008.
- KERMACK, W. O.; MCKENDRICK, A. G. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, v. 115, n. 772, p. 700–721, 1927.
- _____. A contribution to the mathematical theory of epidemics. II. The problem of endemicity. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, v. 138, n. 834, p. 55–83, 1932.
- _____. A contribution to the mathematical theory of epidemics. III. Further studies of the problem of endemicity. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, v. 141, n. 843, p. 94–122, 1933.
- KRÄMER, A.; KRETZSCHMAR, M.; KRICKEBERG, K. (Ed.). *Modern Infectious Disease Epidemiology: Concepts, Methods, Mathematical Models, and Public Health*. New York: Springer, 2010.
- KRICKEBERG, K.; TRONG, P. V.; HANH, P. T. M. *Epidemiology: Key to Public Health*. 2nd. ed. Cham: Springer, 2019.
- KROESE, D. P.; TAIMRE, T.; BOTEV, Z. I. *Handbook of Monte Carlo Methods*. New Jersey: John Wiley & Sons, 2017. ISBN 978-0-470-17793-8.
- LAU, H. et al. The positive impact of lockdown in Wuhan on containing the COVID-19 outbreak in China. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, v. 27, n. 3, p. taaa037, 2020.
- Le Gratiet, L.; MARELLI, S.; SUDRET, B. Metamodel-based sensitivity analysis: Polynomial chaos expansions and gaussian processes. In: *Handbook of Uncertainty Quantification*. Cham: Springer, 2017. p. 1289–1325.
- LI, M. Y. *An Introduction to Mathematical Modeling of Infectious Diseases*. Cham: Springer, 2018.
- LIU, X.; ZHENG, X.; BALACHANDRAN, B. COVID-19: data-driven dynamics, statistical and distributed delay models, and observations. *Nonlinear Dynamics*, v. 101, p. 1527–1543, 2020.
- LYRA, W. et al. COVID-19 pandemics modeling with modified determinist SEIR, social distancing, and age stratification. the effect of vertical confinement and release in Brazil. *PLOS ONE*, v. 15, n. 9, p. e0237627, 2020.
- MARELLI, S. et al. *UQLab user manual – Sensitivity analysis*. [S.l.], 2018. Report # UQLab-V1.1-106.
- MARELLI, S.; SUDRET, B. *UQLab user manual – Polynomial chaos expansions*. [S.l.], 2018. Report # UQLab-V1.1-104.

- MARTCHEVA, M. *An Introduction to Mathematical Epidemiology*. New York: Springer, 2015.
- MERRILL, R. M. *Introduction to epidemiology*. 7th. ed. Massachusetts: Jones and Bartlett Learning, 2017.
- MORRISON, R.; CUNHA JR, A. Embedded model discrepancy: A case study of Zika modeling. *Chaos*, v. 30, p. 051103, 2020.
- MOTULSKY, H. J.; CHRISTOPOULOS, A. *Fitting models to biological data using linear and nonlinear regression: A practical guide to curve fitting*. 2nd. ed. San Diego: GraphPad Software Inc, 2003.
- NELSON, K. E.; WILLIAMS, C. M. *Introduction to epidemiology*. 3rd. ed. Massachusetts: Jones and Bartlett Learning, 2014.
- OKUONGHAE, D.; OMAME, A. Analysis of a mathematical model for COVID-19 population dynamics in Lagos, Nigeria. *Chaos, Solitons and Fractals*, v. 139, p. 110032, 2020.
- PADMANABHAN, P.; SESHAIYER, P. Computational and mathematical methods to estimate the basic reproduction number and final size for single-stage and multistage progression disease models for zika with preventative measures. *Computational and Mathematical Methods in Medicine*, v. 2017, p. 4290825, 2017.
- PAVLACK, B. et al. *EPIDEMIC - Epidemiology Educational Code*. [S.l.]: GitHub, 2021. <www.EpidemicCode.org>.
- PETTERSSON, M. P.; IACCARINO, G.; NORDSTRÖM, J. *Polynomial Chaos Methods for Hyperbolic Partial Differential Equations*. Cham: Springer, 2015.
- PFEFFERBAUM, B.; NORTH, C. S. Mental health and the covid-19 pandemic. *New England Journal of Medicine*, v. 383, p. 510–512, 2020.
- PLATTO, S. et al. History of the COVID-19 pandemic: Origin, explosion, worldwides preading. *Biochemical and Biophysical Research Communications*, v. 0006-291X, 2020.
- RICHARDS, F. J. A flexible growth function for empirical use. *Journal of Experimental Botany*, v. 10, n. 29, p. 290–300, 1959.
- ROCK, K. et al. Dynamics of infectious diseases. *Reports on Progress in Physics*, v. 77, n. 2, p. 026602, 2014.
- RUBINSTEIN, R. Y.; KROESE, D. P. *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation, and Machine Learning*. New York: Springer, 2004.
- SALTELLI, A. et al. Why so many published sensitivity analyses are false: A systematic review of sensitivity analysis practices. *Environmental Modelling and Software*, v. 114, p. 29–39, 2019.
- _____. *Global Sensitivity Analysis. The Primer*. West Sussex: John Wiley and Sons, 2008.

- _____. *Sensitivity Analysis in Practice: A guide to assessing scientific models*. West Sussex: John Wiley and Sons, 2004.
- SCHWARTZ, R. A.; KAPILA, R. Pandemics throughout the centuries. *Clinics in dermatology*, v. 39, n. 1, p. 5–8, 2021.
- SHEN, C. Y. Logistic growth modelling of COVID-19 proliferation in China and its international implications. *International Journal of Infectious Diseases*, v. 96, p. 582–589, 2020.
- SMITH, C. R.; ERICKSON, G. J.; NEUDORFER, P. O. *Maximum Entropy and Bayesian Methods*. Dordrecht: Springer, 1991.
- SMITH, D. L. et al. Ross, Macdonald, and a Theory for the Dynamics and Control of Mosquito-Transmitted Pathogens. *PLOS Pathogens*, v. 8, n. 4, p. e1002588, 2012.
- SMITH, R. C. *Uncertainty Quantification: Theory, Implementation, and Applications*. Philadelphia: SIAM, 2014.
- SOBOL, I. M. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation*, v. 55, n. 1-3, p. 271–280, 2001.
- SOIZE, C. *Uncertainty Quantification: An Accelerated Course with Advanced Applications in Computational Engineering*. Cham: Springer, 2017.
- SZKLO, M.; NIETO, F. J. *Epidemiology: Beyond the basics*. 4th. ed. Burlington: Jones & Bartlett Learning, 2019.
- TANG, X. et al. On the origin and continuing evolution of SARS-CoV-2. *National Science Review*, v. 7, n. 6, p. 1012–1023, 2020.
- TARANTOLA, A. *Inverse Problem Theory*. Philadelphia: SIAM, 2005.
- TOSIN, M.; CÔRTEZ, A. M. A.; CUNHA JR, A. A tutorial on sobol' global sensitivity analysis applied to biological models. In: *Networks in Systems Biology*. Cham: Springer, 2020. p. 93–118.
- TOSIN, M.; DANTAS, E.; CUNHA JR., A. *ZikaVD - Zika Virus Dynamics*. [S.l.]: GitHub, 2021. <<https://github.com/americocunhajr/ZikaVD>>.
- UDWADIA, F. E. Some results on maximum entropy distributions for parameters known to lie in finite intervals. *SIAM Review*, v. 31, n. 1, p. 103–109, 1989.
- VALENTINE, G.; MARQUEZ, L.; PAMMI, M. Zika virus-associated microcephaly and eye lesions in the newborn. *Journal of the Pediatric Infectious Diseases Society*, v. 5, n. 3, p. 323–328, 2016.
- VASCONCELOS, G. L. et al. Standard and anomalous waves of COVID-19: A multiple-wave growth model for epidemics. MedRxiv 2021.01.31.21250867, v4. 2021.
- _____. Situation of COVID-19 in Brazil: An analysis via growth models as implemented in the ModInterv system for monitoring the pandemic. MedRxiv 2021.03.29.21254542, v2. 2021.

_____. Power law behaviour in the saturation regime of fatality curves of the COVID-19 pandemic. *Scientific Reports*, v. 11, n. 4619, 2021.

VERHULST, P.-F. Notice sur la loi que la population poursuit dans son accroissement. *Correspondance Mathématique et Physique*, v. 10, p. 113–121, 1938.

_____. Recherches mathématiques sur la loi d'accroissement de la population. *Nouveaux Memoires de l'Academie Royale des Sciences et Belles-Lettres de Bruxelles*, v. 18, p. 1–41, 1945.

_____. Deuxième mémoire sur la loi d'accroissement de la population. *Mémoires de l'Académie Royale des Sciences, des Lettres et des Beaux-Arts de Belgique*, v. 20, p. 142–173, 1947.

VYNNYCKY, E.; WHITE, R. G. *An introduction to infectious disease modelling*. Oxford: Oxford University Press, 2010.

WASSERMAN, L. *All of Statistics: A Concise Course in Statistical Inference*. New York: Springer, 2004.

WATKINS, K. Emerging infectious diseases: a review. *Current Emergency and Hospital Medicine Reports*, v. 6, n. 3, p. 86–93, 2018.

WIRATSUDAKUL, A.; SUPARIT, P.; MODCHANG, C. Dynamics of Zika virus outbreaks: an overview of mathematical modeling approaches. *PeerJ*, v. 6, p. e4526, 2018.

WORLD HEALTH ORGANIZATION. *Zika situation report: Zika virus, Microcephaly and Guillain-Barré syndrome*. 2016. https://apps.who.int/iris/bitstream/handle/10665/204491/zikasitrep_26Feb2016_eng.pdf?sequence=1.

_____. *WHO SAGE values framework for the allocation and prioritization of COVID-19 vaccination*. 2020. https://apps.who.int/iris/bitstream/handle/10665/334299/WHO-2019-nCoV-SAGE_Framework-Allocation_and_prioritization-2020.1-eng.pdf.

_____. *COVID 19 public health emergency of international concern (PHEIC), global research and innovation forum: Towards a research roadmap*. 2020. [https://www.who.int/publications/m/item/covid-19-public-health-emergency-of-international-concern-\(pheic\)-global-research-and-innovation-forum](https://www.who.int/publications/m/item/covid-19-public-health-emergency-of-international-concern-(pheic)-global-research-and-innovation-forum).

_____. *EPI-WIN, Update 36: What we know about Long-term effects of COVID-19*. 2020. https://www.who.int/docs/default-source/coronaviruse/risk-comms-updates/update-36-long-term-symptoms.pdf?sfvrsn=5d3789a6_5. Accessed: 02/26/2021.

_____. *MERS situation update, January 2020*. 2020. <https://applications.emro.who.int/docs/EMCSR254E.pdf?ua=1>.

_____. *Timeline: WHO's COVID-19 response*. 2020. <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/interactive-timeline#event-0>. Accessed: 02/26/2021.

_____. *Status of COVID-19 vaccines within WHO EUL/PQ evaluation process (20 January 2021)*. 2021. https://extranet.who.int/pqweb/sites/default/files/documents/Status_COVID_VAX_20Jan2021_v2.pdf. Accessed: 01/25/2021.

_____. *WHO coronavirus disease (COVID-19) dashboard*. 2021. (<https://covid19.who.int/>).

XIONG, J. et al. Impact of COVID-19 pandemic on mental health in the general population: A systematic review. *Journal of Affective Disorders*, v. 227, p. 55–64, 2020.

XIU, D. *Numerical Methods for Stochastic Computations: A Spectral Approach*. New Jersey: Princeton University Press, 2010.

XIU, D.; KARNIADAKIS, G. E. The wiener–askey polynomial chaos for stochastic differential equations. *SIAM Journal on Scientific Computing*, v. 24, n. 2, p. 619–644, 2002.

ZOU, Y. et al. Outbreak analysis with a logistic growth model shows COVID-19 suppression dynamics in china. *PLOS ONE*, v. 15, n. 6, p. e0235247, 2020.

APPENDIX A – Sensitivity analysis for model selection

In the two previous results chapters, the main goal was to applied the UQ framework. Even so, the use of sensitivity analysis as source of information deserves a special attention. In this appendix is presented a model selection method using Sobol indices results.

A.1 Joint Sobol indices

Let the random variable $Y = \mathcal{M}(\mathbf{X})$, with $\mathbf{X} \sim f_{\mathbf{X}}$, and its total Sobol indices S_i^T , constructed as presented in the Section 3.4. Normally, the intention with these is just detect the parameters that satisfy the higher indices. However, this neglect the contrast between the parameters contributions. For example, in a situation of two selected inputs X_1 and X_2 , the total index for both be closer one to another (as the both ≈ 0.4), or maybe X_1 concentrates the density of the total indices. Then, a novel joint Sobol index is constructed as follows

$$S^J = \prod_{i=1}^n S_i^T. \quad (52)$$

While the total indices measures the full contribution given for each parameter, the idea is to use the joint index as reference of how that contribution as distributed between the inputs. If one parameter have a total index too close to 1, the other will have very small indices, and the joint index will be more closer to zero. This measure increases only if the density of total contribution is balanced between the inputs. Of course, the joint index naturally tends to be a small number because of the parameters for which the contribution for the total variance is tiny. To avoid work with to small values, the log-joint index is given by

$$LS^J = \sum_{i=1}^n \log S_i^T. \quad (53)$$

Now, the distribution of the total variance is as equal as the log-index is closer to zero.

A.2 Sobol indices-based model selection

When start studying a phenomenon, the choice of the reference model is extremely important. Normally, real data are used to calibrate some candidate models and help decide between them. If those models have similar results, the simplest is probably be recommended. But if there is no trustful data to be used, the decision of which model to use becomes way more complicated. In that sense, a decision criterion could be to favor the one who the parameters contributions is balances, or the opposite. This last makes more sense because implies that a minor set of most important parameters can be selected. Have less parameters to work with in future uncertainty quantification scenarios is desirable. Therefore, assuming a set of model candidates $M = \{\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_{N_c}\}$, will be selected the one who satisfies

$$\mathcal{M}^* = \operatorname{argmax}_{\mathcal{M}_j \in M} LS^J(\mathcal{M}_j). \quad (54)$$

Although, this measure do not cover the scenarios of time dependence Sobol indices. Also, it is necessary to include a penalty based in the number of inputs of the model candidate. Thus, the Akaike information criterion (AKAIKE, 1974) is applied using the log-indices as likelihood. Therefore, each model will be classified through an index calculated as follows

$$AIC = 2n - 2LS^J, \quad (55)$$

where n is the number of inputs and LS_i^J is the log-joint Sobol index. When working with time dependence model QoIs, the criterion is adapted to the form

$$AIC = \frac{2nd}{d-n-1} - 2\hat{L}, \quad (56)$$

with

$$\hat{L} = \operatorname{argmax}_{t_j \in \mathcal{T}} LS^J(t_j). \quad (57)$$

Here $t_j \in \mathcal{T}$ are the time instants to which the QoI is defined. The selected model will be the one who satisfies lower classifier value.

A.3 Study case

To illustrate the practical use of the methodology, the Sobol-Akaike criterion will be applied into a compartmental model used to model COVID-19 in Nigeria (OKUONG-HAE; OMAME, 2020), but adapting for the Rio de Janeiro (city) population. It consider a set of 7 compartments: Susceptible (S); Exposed (E); Asymptomatic infectious (A); Symptomatic infectious (I); Tested (T); Recovered (R); Deceased (D). The model is governed by the following set of equations

$$\begin{aligned}
\frac{dS}{dt} &= -S \frac{(\beta_A A + \beta_I I)}{N - T - D} , \\
\frac{dE}{dt} &= S \frac{(\beta_A A + \beta_I I)}{N - T - D} - \alpha E , \\
\frac{dA}{dt} &= \nu \alpha E - (\gamma_A + \theta_A) A , \\
\frac{dI}{dt} &= (1 - \nu) \alpha E - (\gamma_I + \theta_I + \mu_I) I , \\
\frac{dT}{dt} &= \theta_A A + \theta_I I - (\gamma_T + \mu_T) T , \\
\frac{dR}{dt} &= \gamma_A A + \gamma_I I + \gamma_T T , , \\
\frac{dD}{dt} &= \mu_I I + \mu_T T ,
\end{aligned} \tag{58}$$

$N = 6.7 \times 10^6$ is the original total population, β represents the transmission rates, $1/\alpha$ is the incubation period, $1/\gamma$ indicates the recovery periods, ν is the asymptomatic fraction, θ is referred to the testing rates, and μ are the mortality rates. If assuming The model quantity of interest is the daily number of cumulative deaths, $D(t)$. From that original model, a set of candidates models is construct as follows

$$\left\{ \begin{array}{l} \text{SEAITRD base model .} \\ \text{inputs: } \beta_A, \beta_I, \alpha, \nu, \gamma_A, \gamma_I, \gamma_T, \theta_A, \theta_I, \mu_I, \mu_T ; \end{array} \right. \tag{Model 1}$$

$$\left\{ \begin{array}{l} \text{SEAITRD model adapted with } \gamma_A = \gamma_I . \\ \text{inputs: } \beta_A, \beta_I, \alpha, \nu, \gamma_A, \gamma_T, \theta_A, \theta_I, \mu_I, \mu_T ; \end{array} \right. \tag{Model 2}$$

$$\left\{ \begin{array}{l} \text{SEAITRD model adapted with } \theta_A = \theta_I . \\ \text{inputs: } \beta_A, \beta_I, \alpha, \nu, \gamma_A, \gamma_I, \gamma_T, \theta_A, \mu_I, \mu_T ; \end{array} \right. \tag{Model 3}$$

$$\left\{ \begin{array}{l} \text{SEAITRD model adapted with } \theta_A = 0 . \\ \text{inputs: } \beta_A, \beta_I, \alpha, \nu, \gamma_A, \gamma_I \gamma_T, \theta_I, \mu_I, \mu_T ; \end{array} \right. \quad (\text{Model 4})$$

$$\left\{ \begin{array}{l} \text{SEAITRD model adapted with } \nu = 0.25 . \\ \text{inputs: } \beta_A, \beta_I, \alpha, \gamma_A, \gamma_I \gamma_T, \theta_A, \theta_I, \mu_I, \mu_T ; \end{array} \right. \quad (\text{Model 5})$$

$$\left\{ \begin{array}{l} \text{SEAITRD model adapted with } \nu = 0.5 . \\ \text{inputs: } \beta_A, \beta_I, \alpha, \gamma_A, \gamma_I \gamma_T, \theta_A, \theta_I, \mu_I, \mu_T . \end{array} \right. \quad (\text{Model 6})$$

To decide between the candidates, without using real data, the a global sensitivity analysis is performed using the parameters supports and initial conditions listed in the Table 7. For each model, 500 samples were used to construct the Sobol indices for the number of deaths in the first 30 days after the initial condition. By obtaining the joint Sobol indices, the AIC classifiers can be calculated for the six model candidates. These values are reunited in the Table 8.

Table 7 - Based SEAITRD model initial conditions and parameters supports.

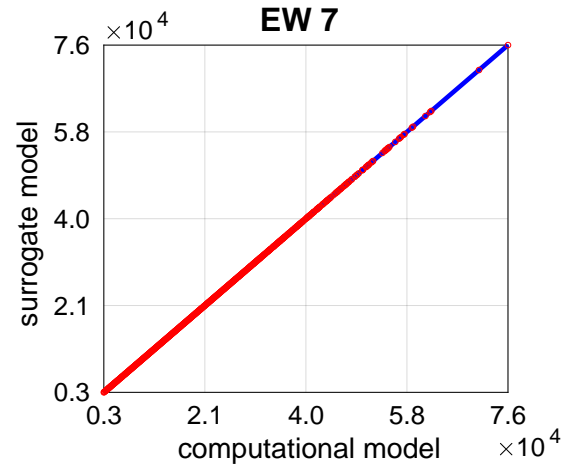
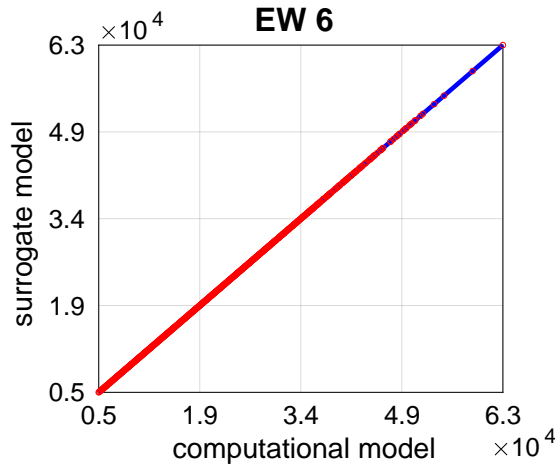
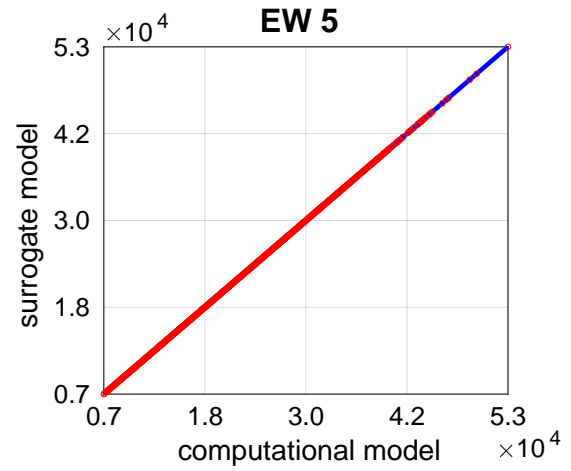
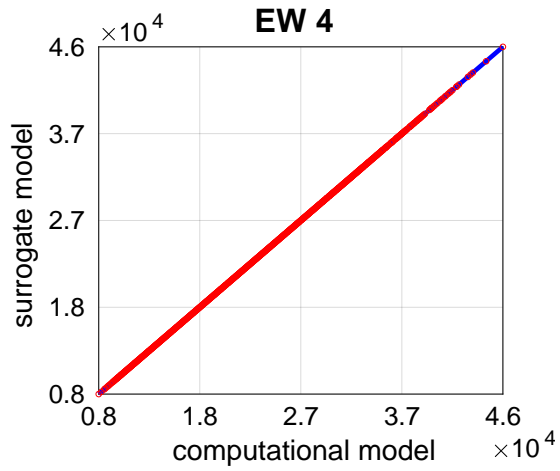
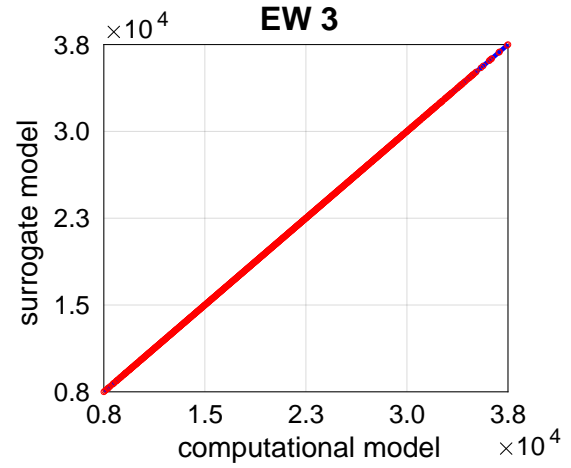
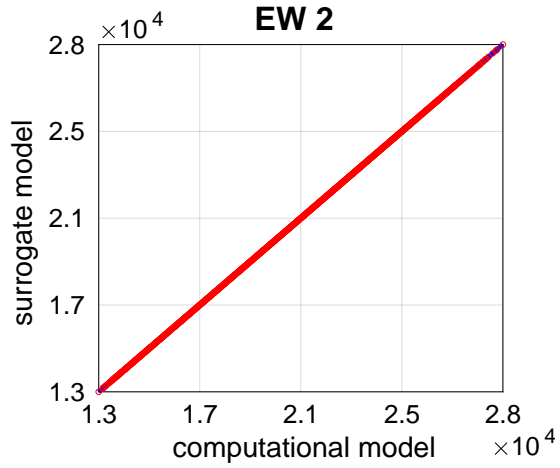
Parameter	β_A	β_I	α	ν	γ_A	γ_I	
Support	[0.25,0.5]	[0.25,0.5]	[0,1]	[0,1]	[1/30,1/3]	[1/30,1/3]	
Parameter	γ_T	θ_A	θ_I	μ_I	μ_T		
Support	[1/30,1/3]	$[10^{-4}, 10^{-3}]$	$[10^{-4}, 10^{-3}]$	[0.001,0.1]	[0.001,0.1]		
Group	S	E	A	I	T	R	D
Initial condition	6699925	25	25	25	0	0	0

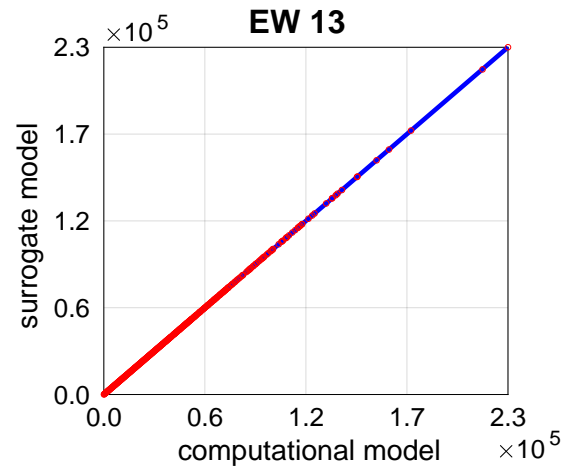
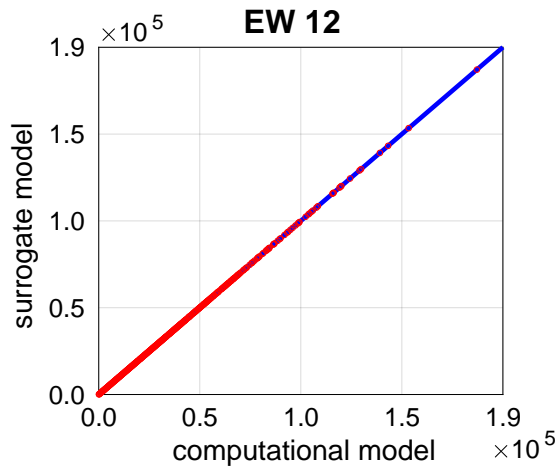
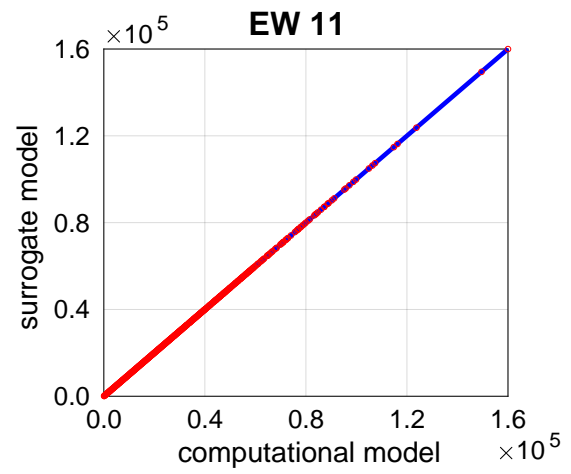
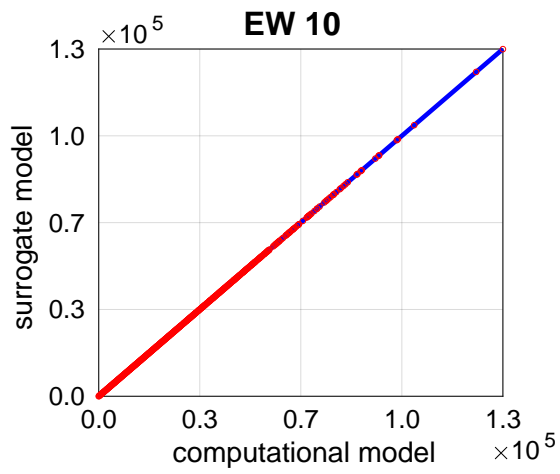
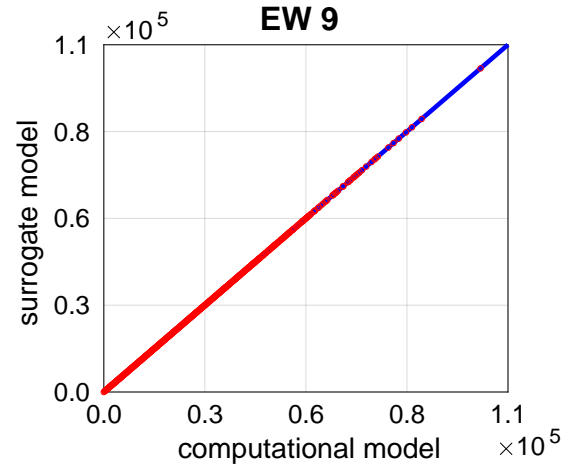
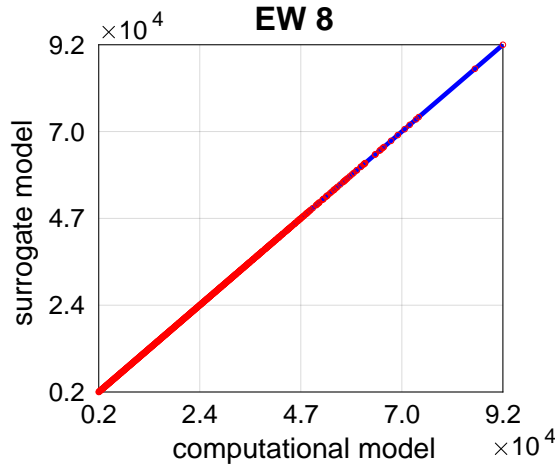
Table 8 - Classification values for each candidate model.

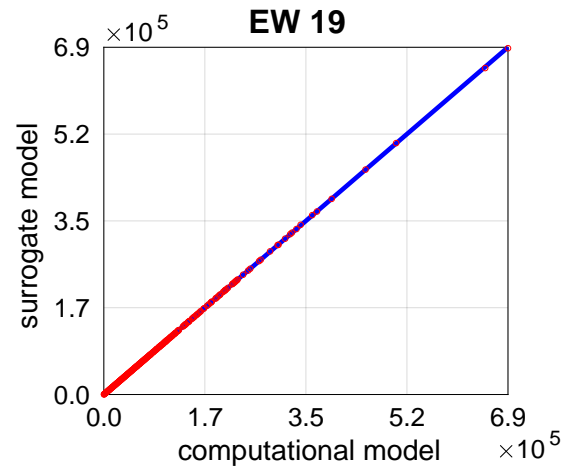
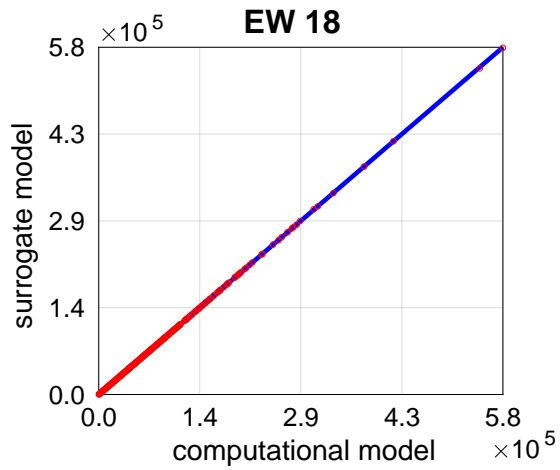
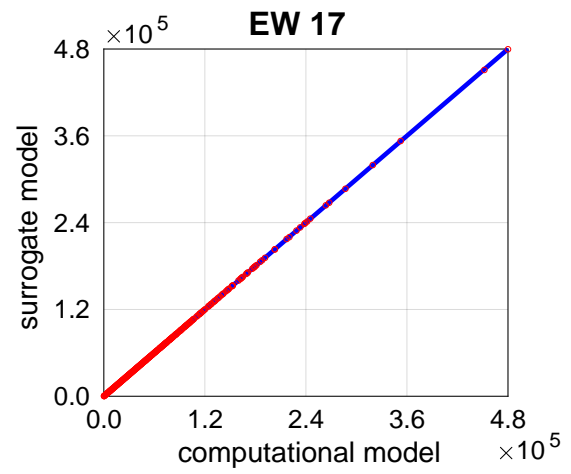
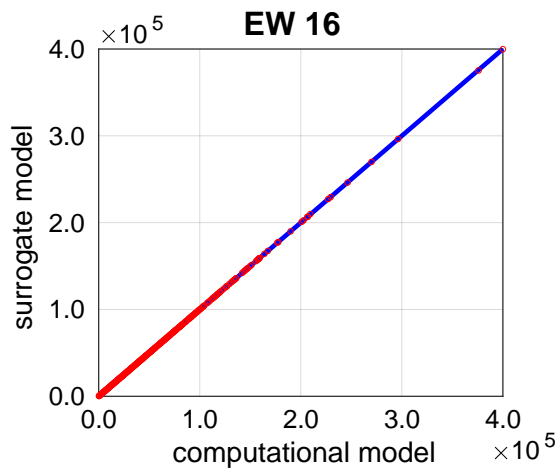
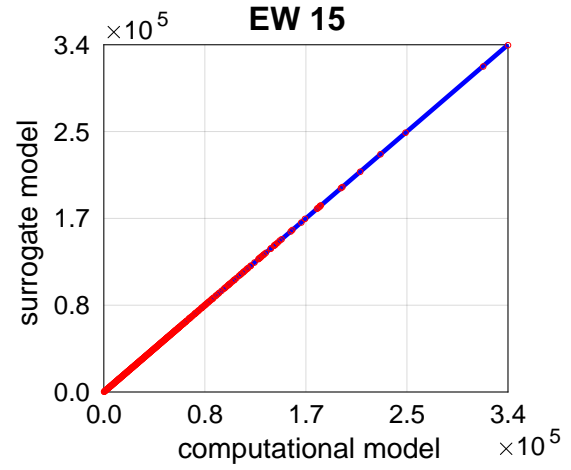
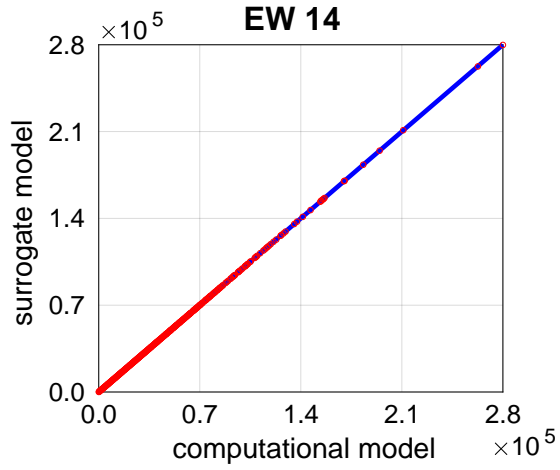
Candidate	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
AIC	82.99	72.12	68.09	68.26	78.50	74.37

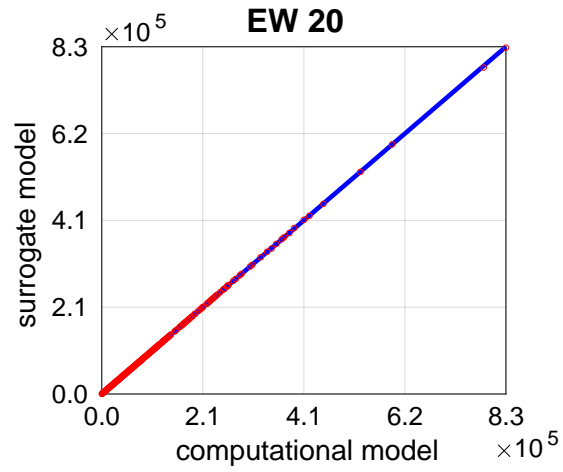
APPENDIX B – Supplementary material

B.1 Chapter 5

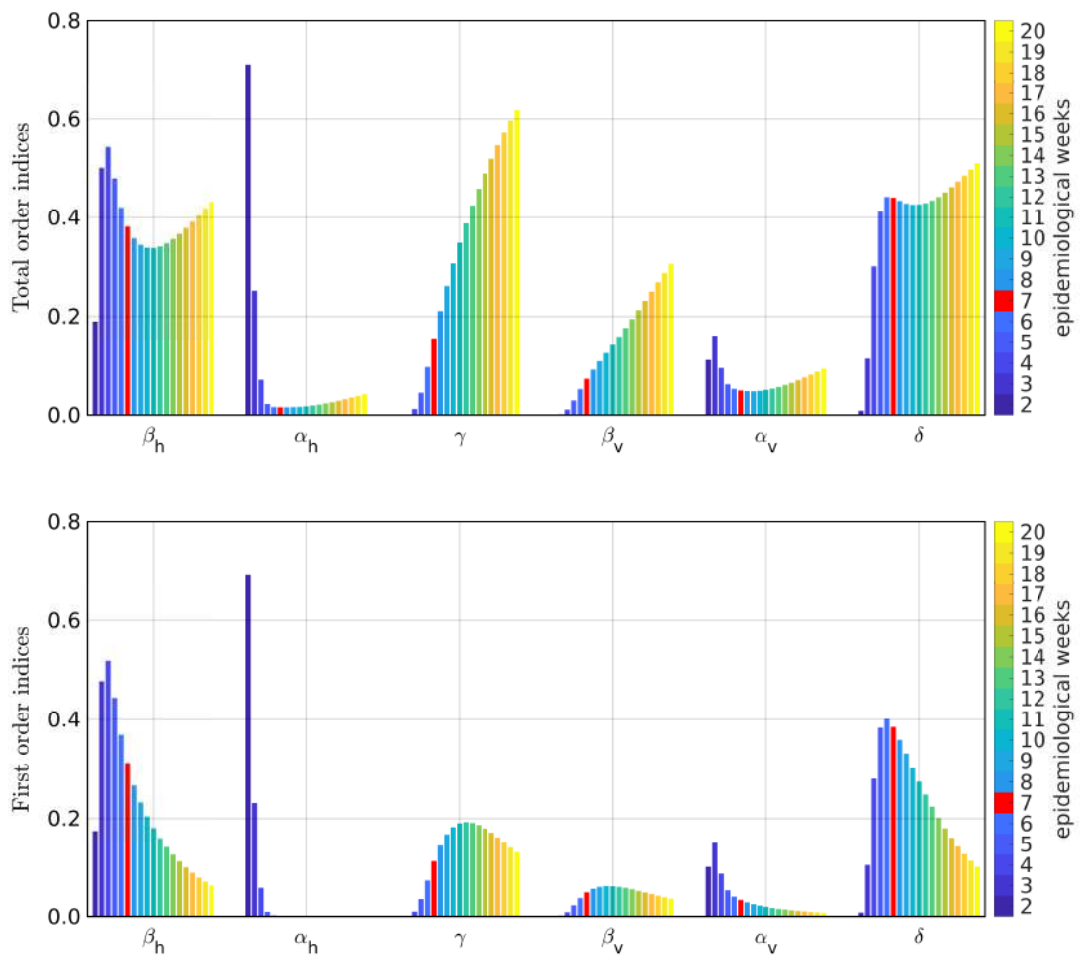
B.1.1 PCE validation curves

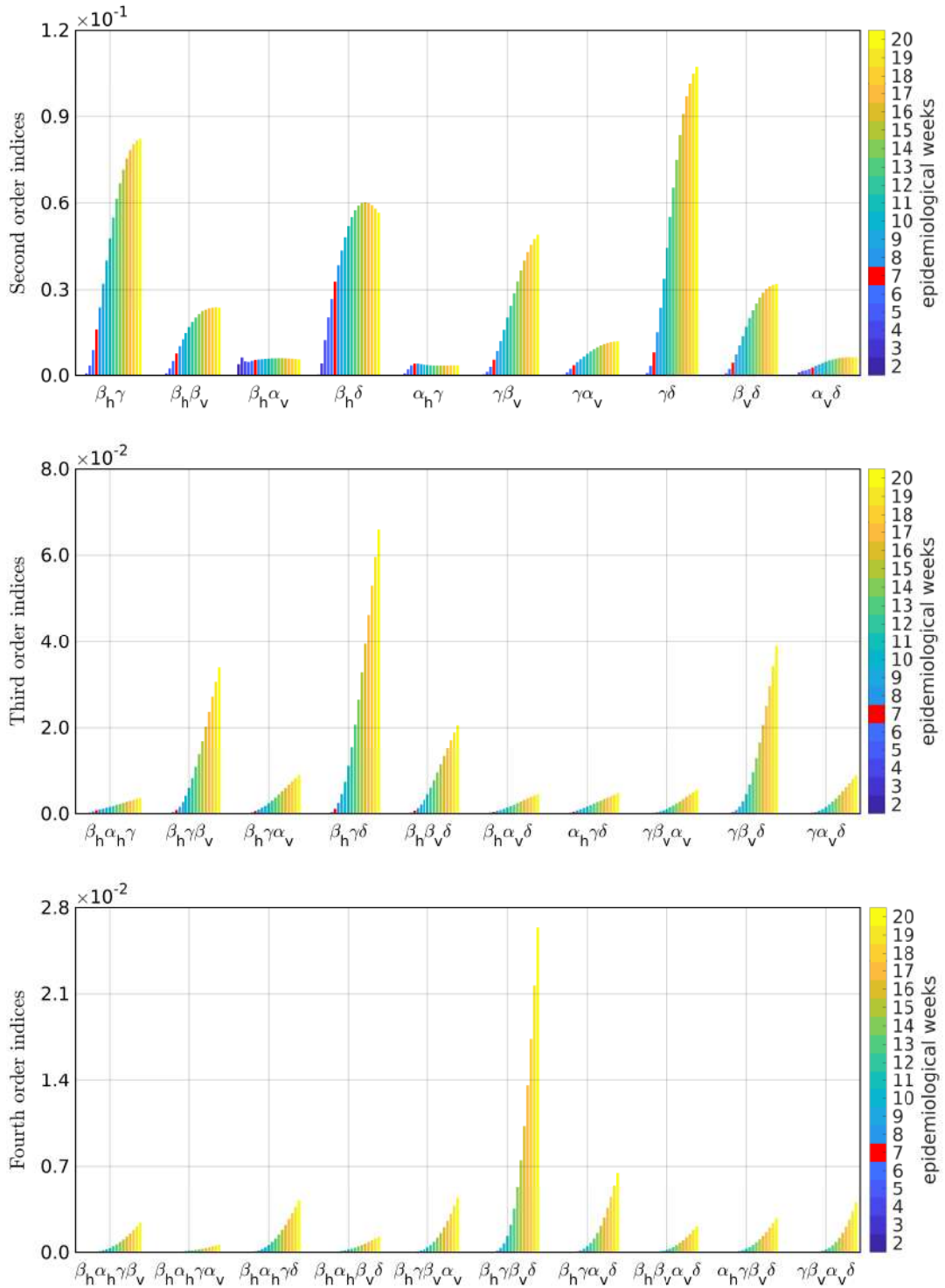


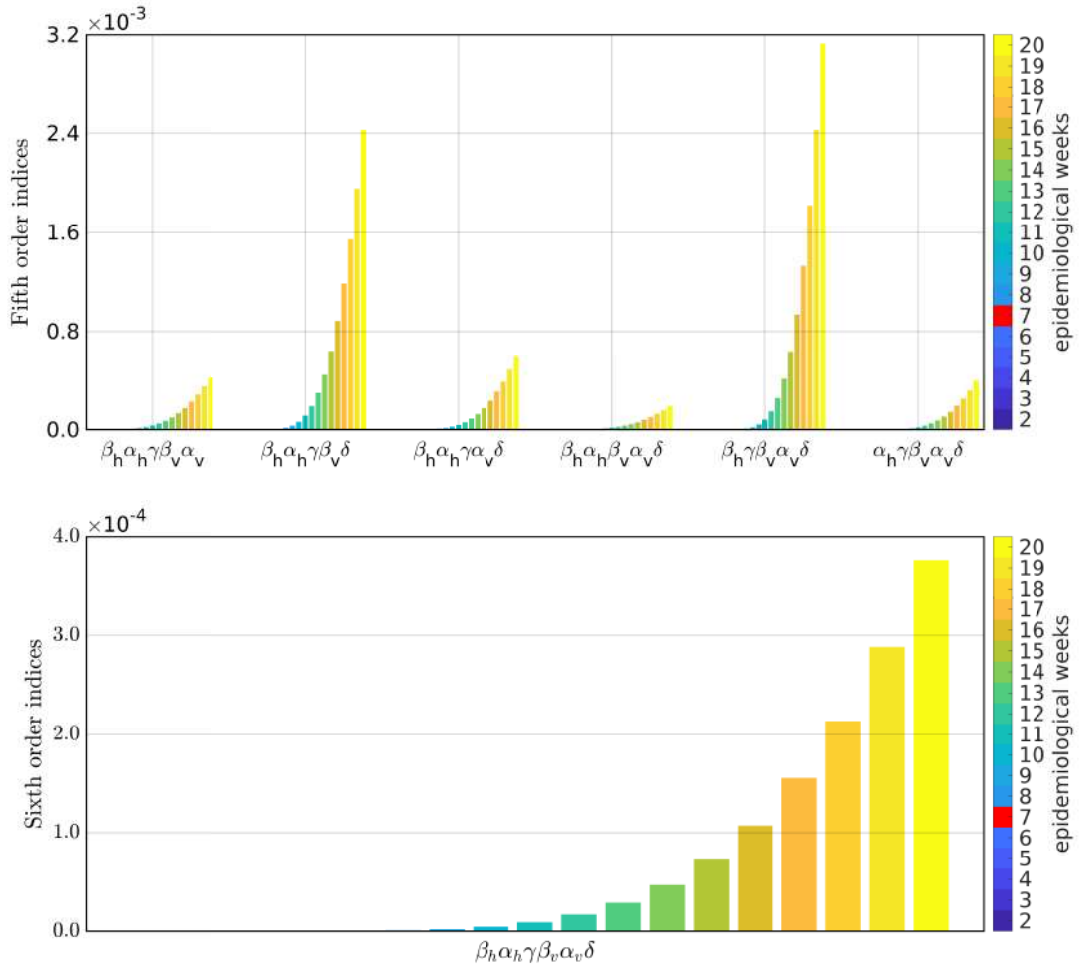




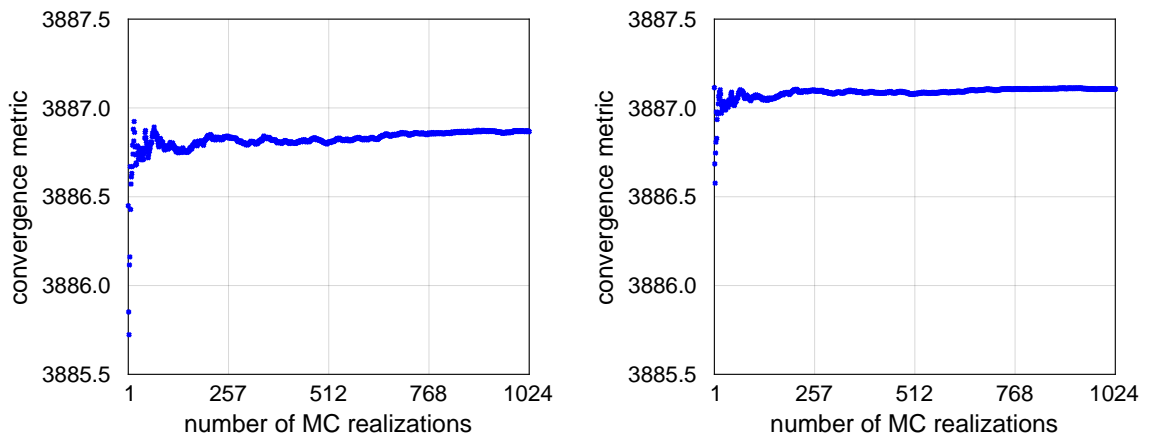
B.1.2 Sobol indices

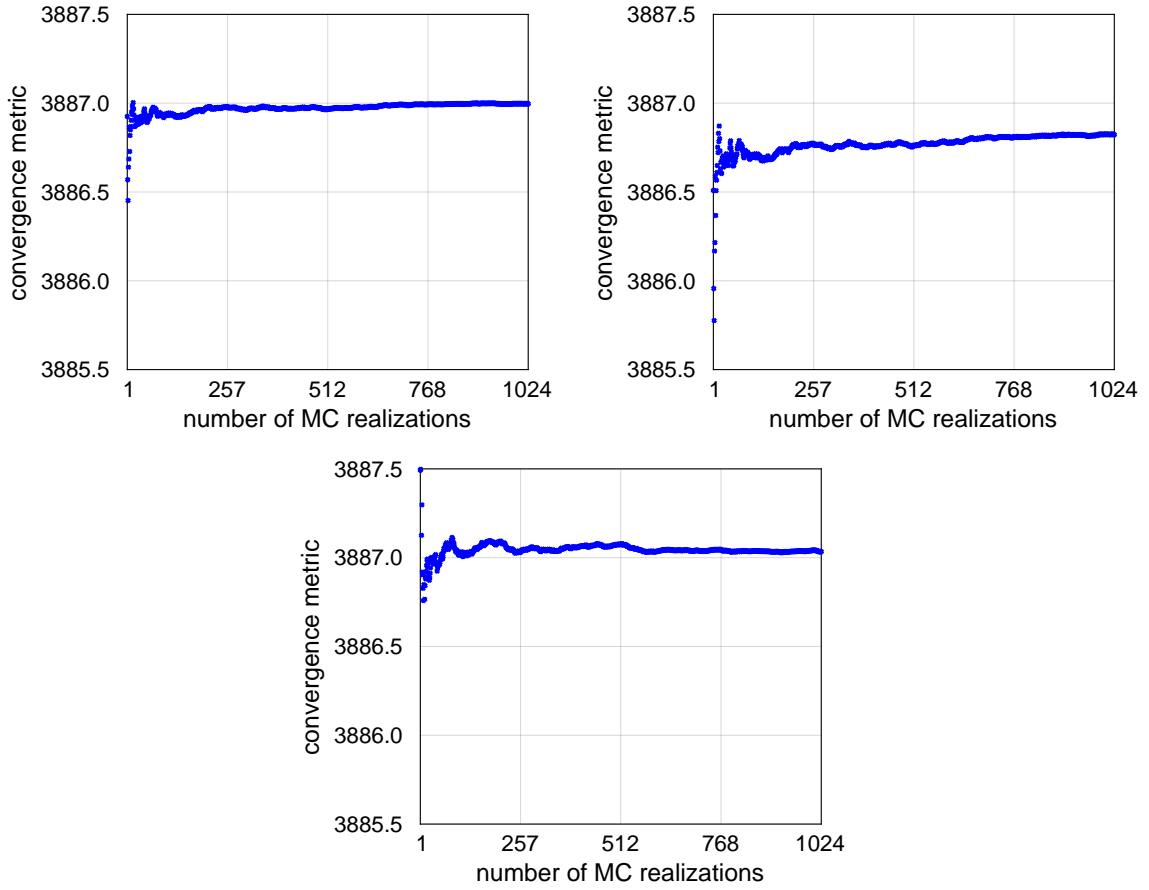






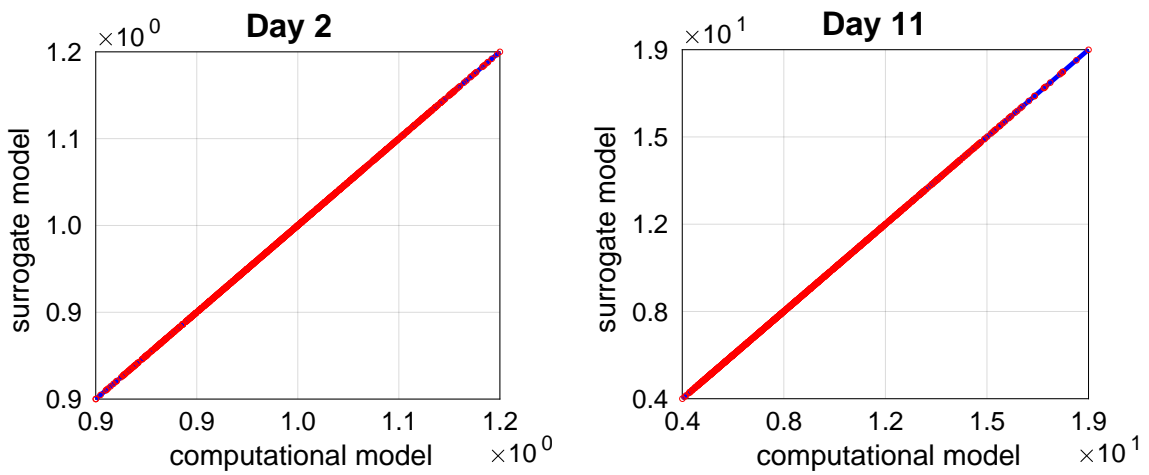
B.1.3 Monte Carlo convergence

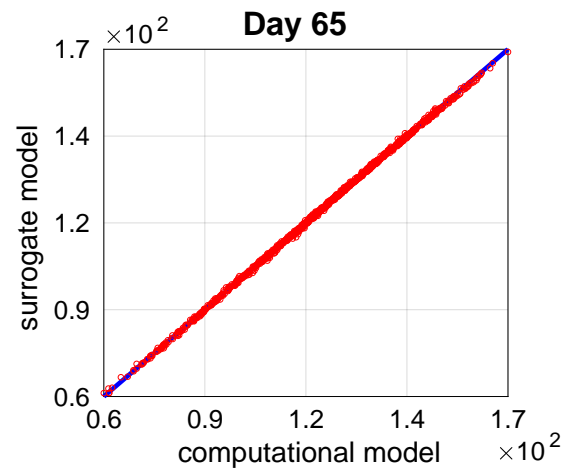
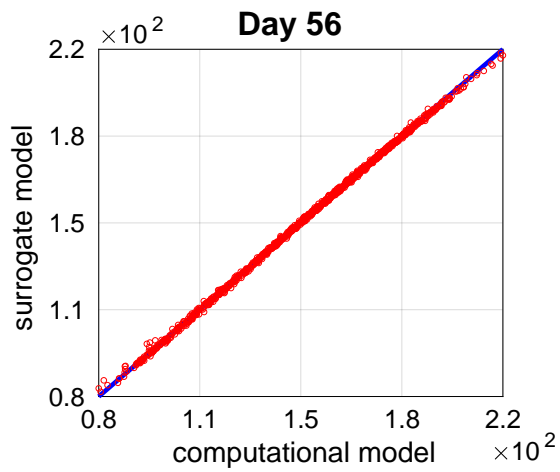
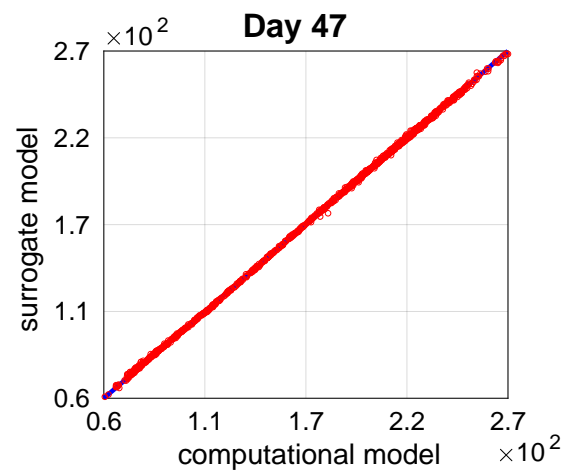
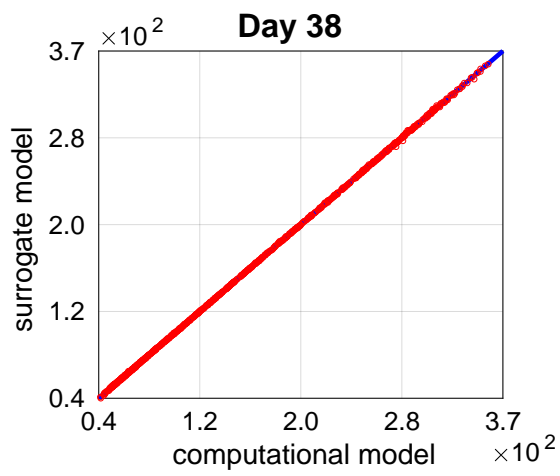
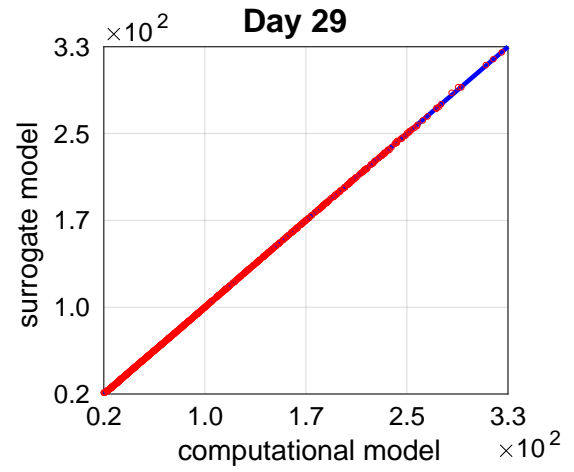
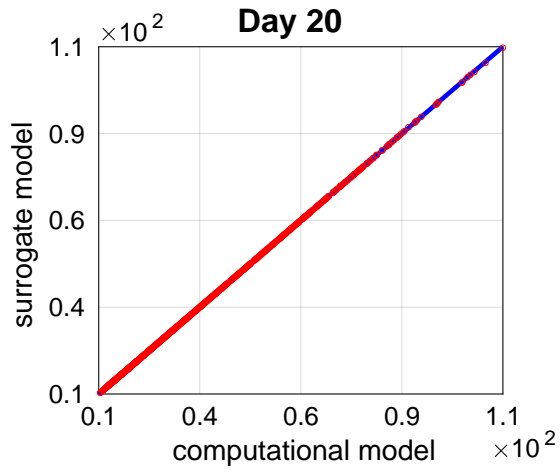


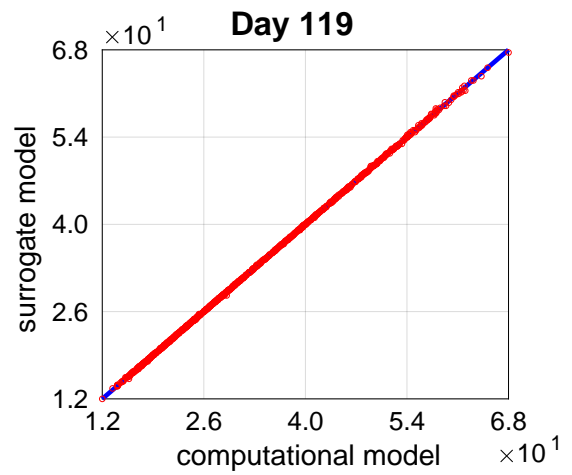
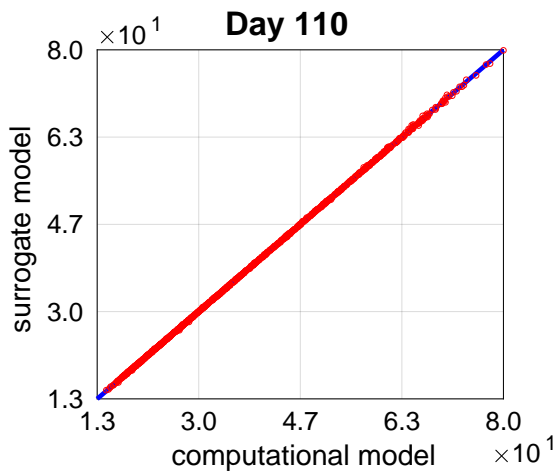
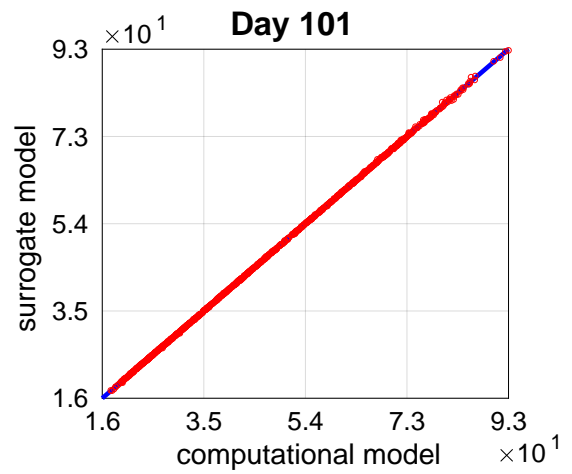
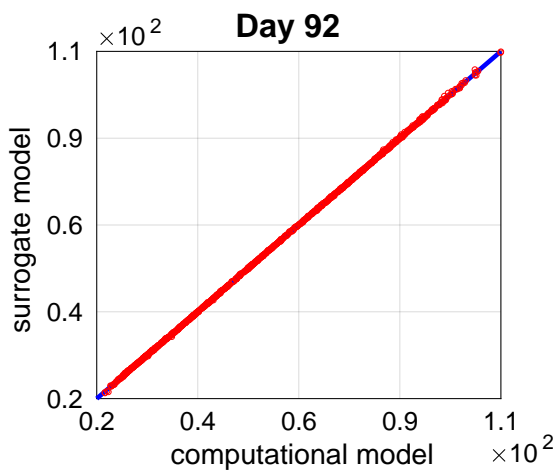
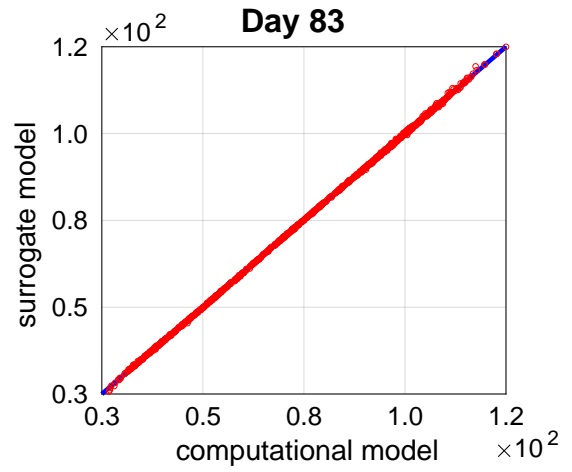
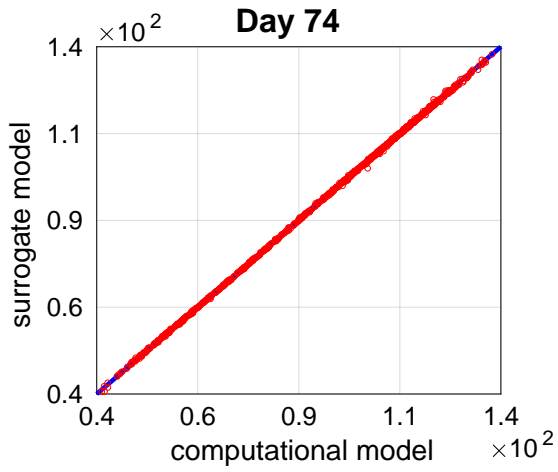


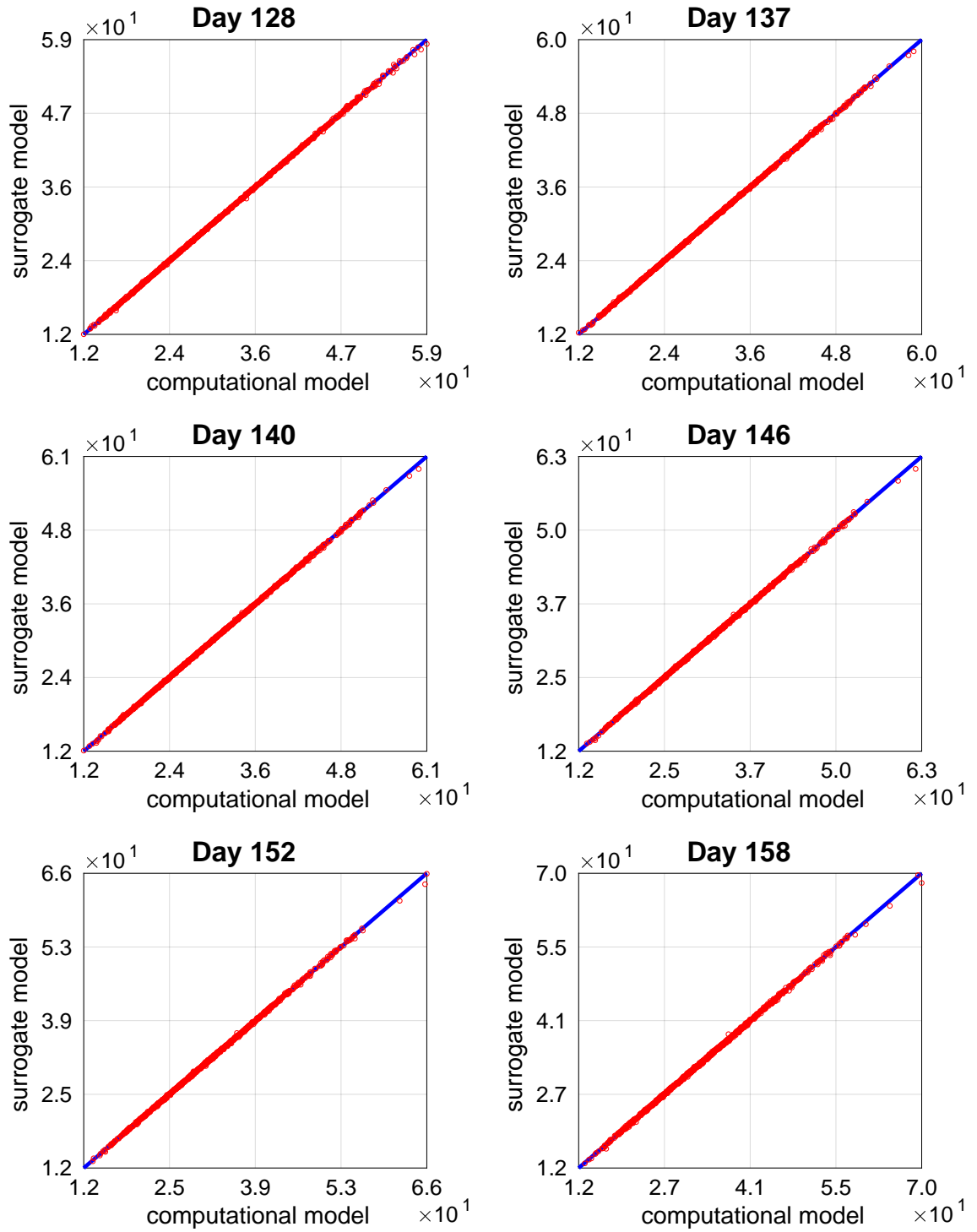
B.2 Chapter 6

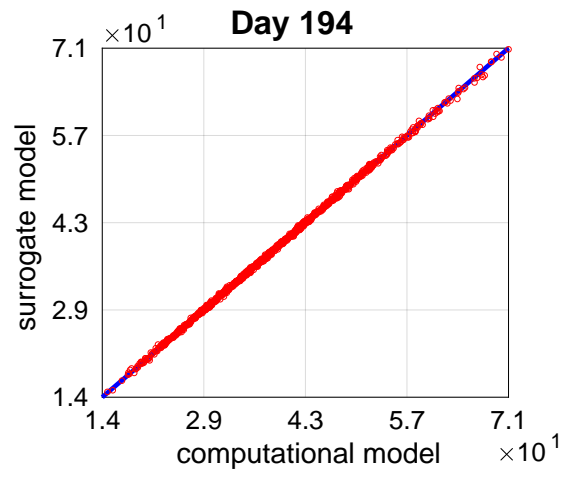
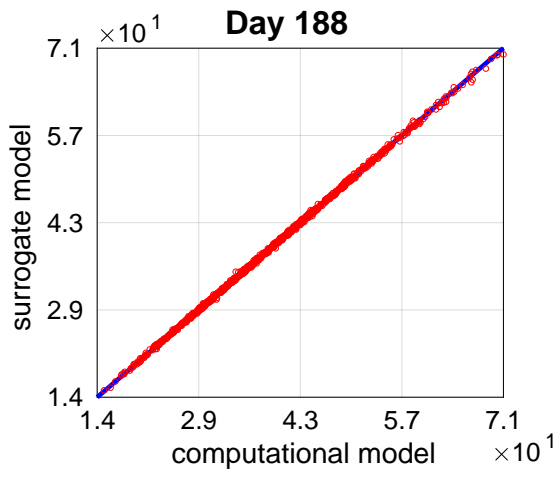
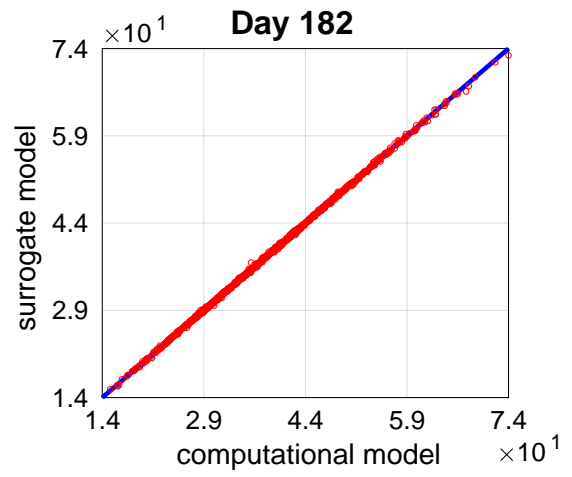
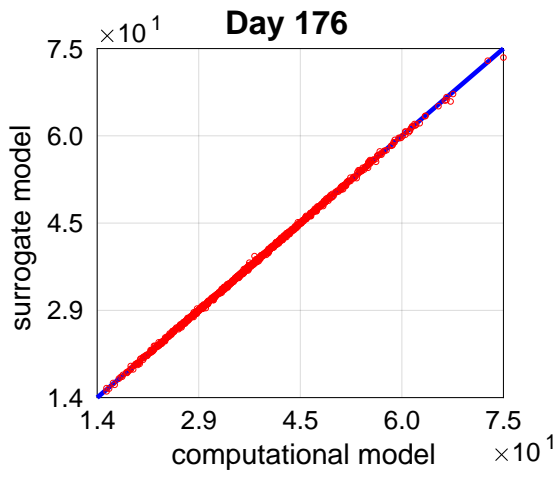
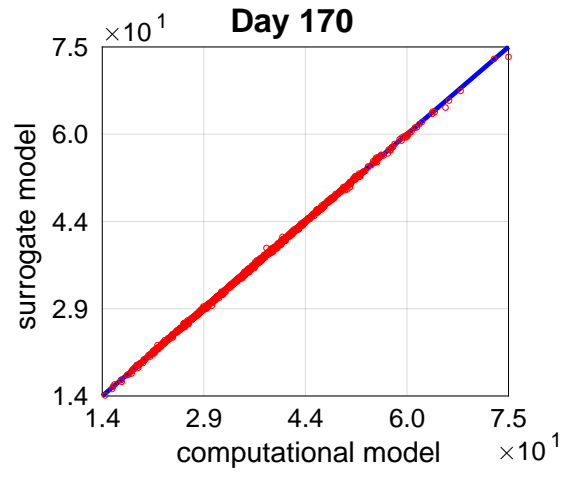
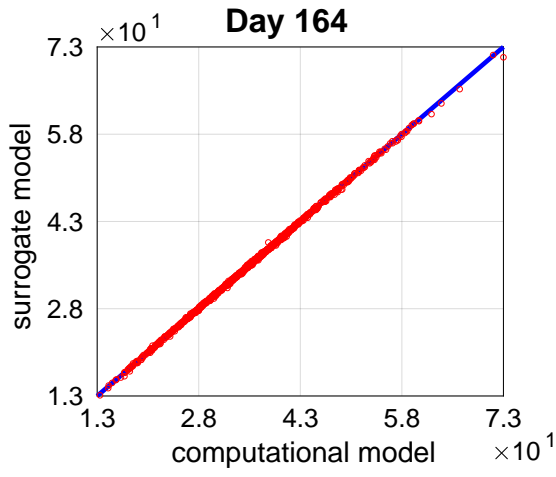
B.2.1 PCE validation curves

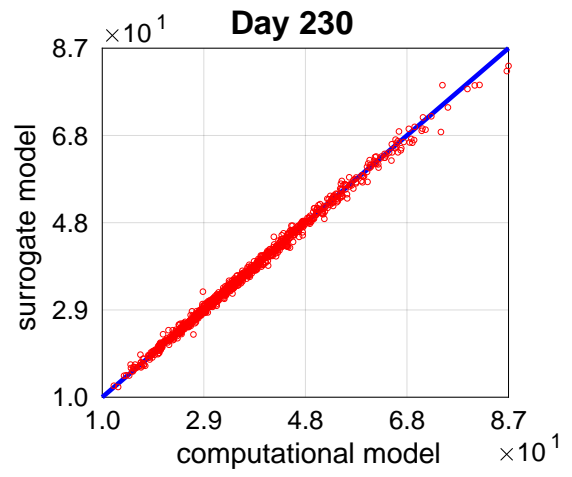
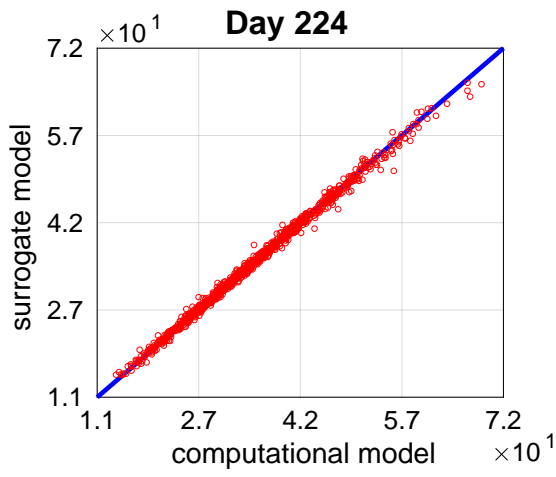
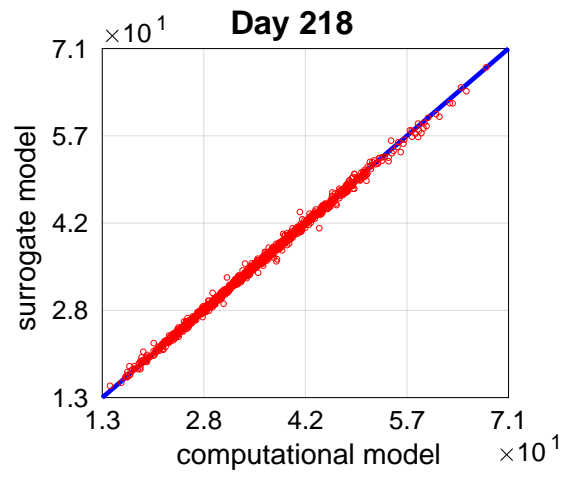
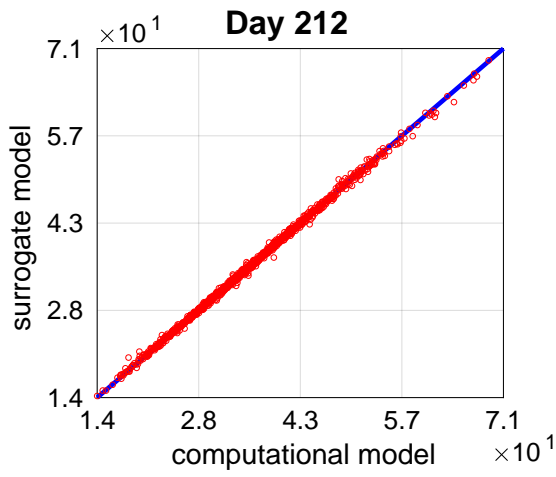
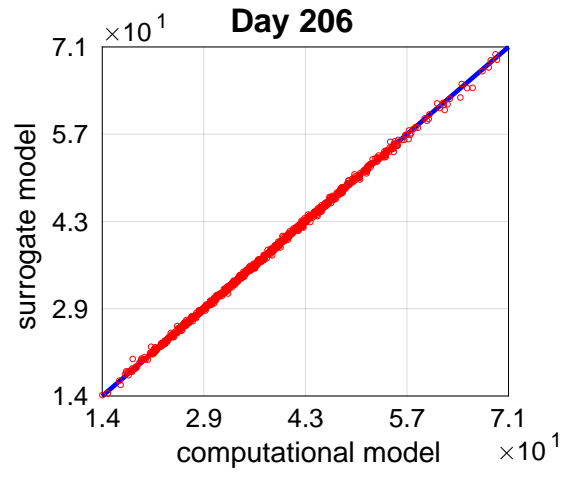
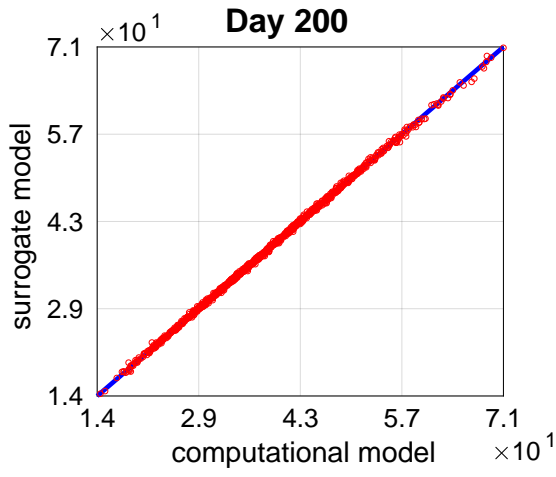


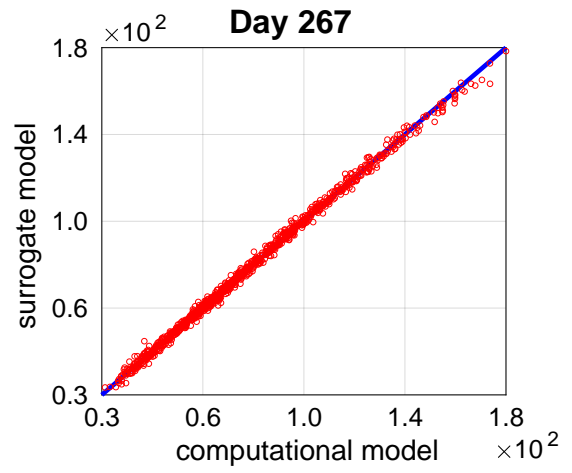
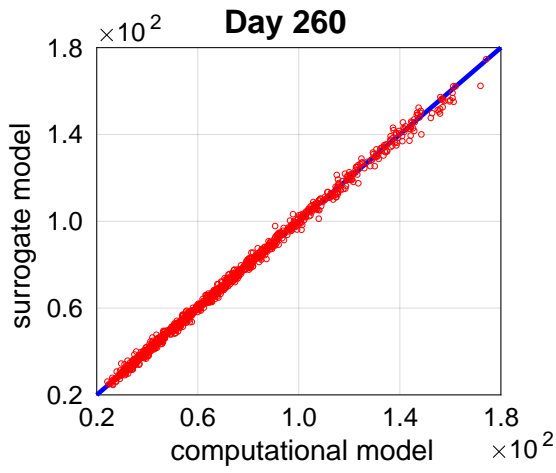
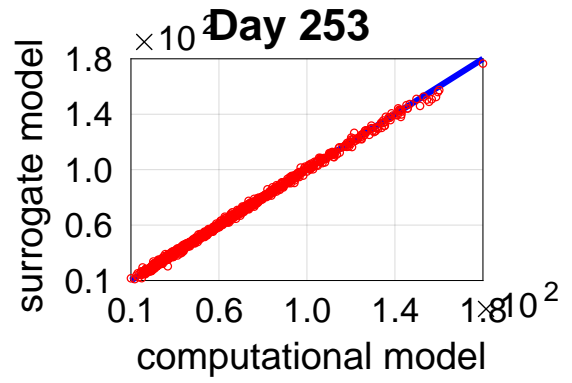
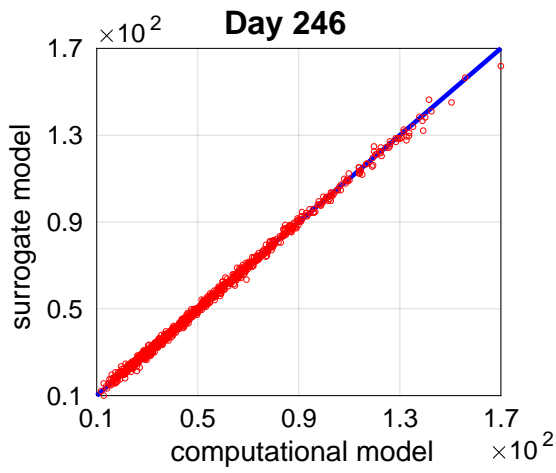
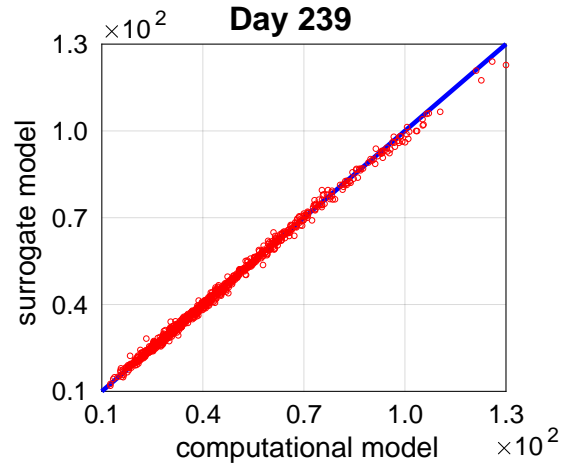
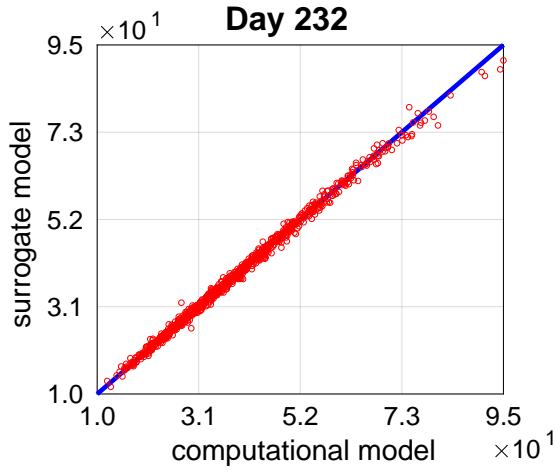


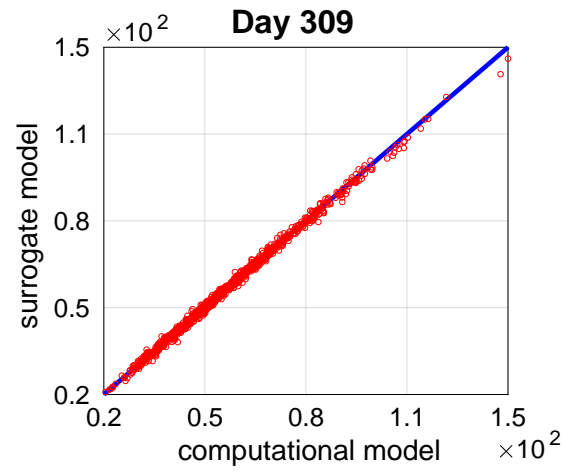
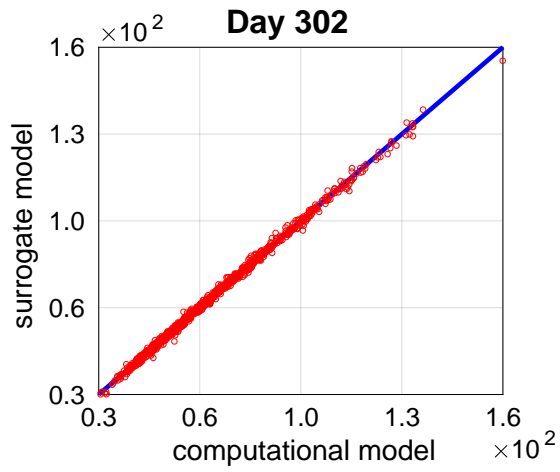
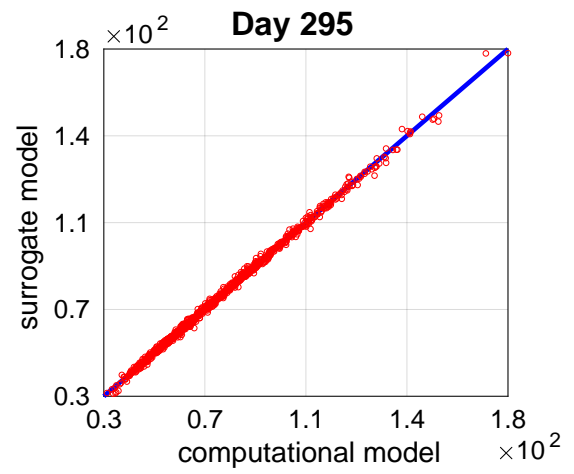
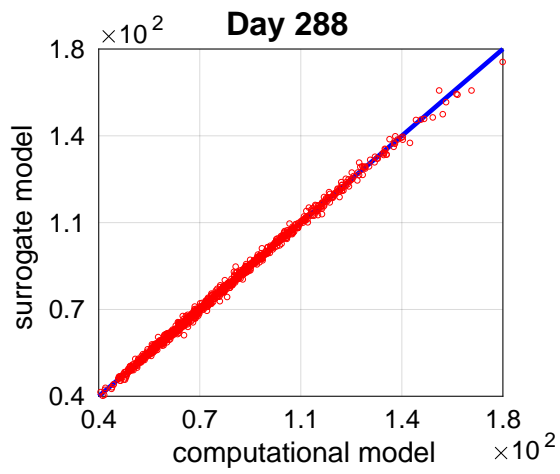
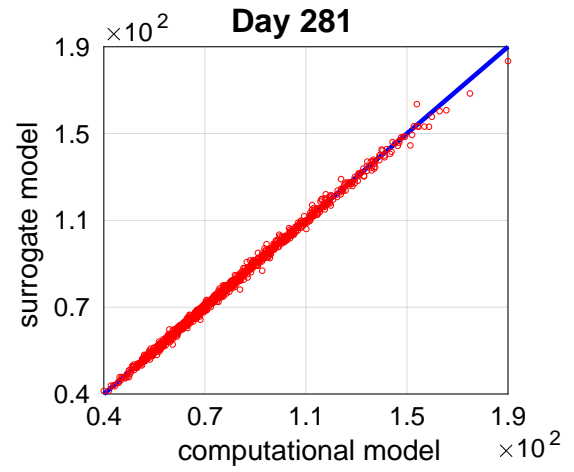
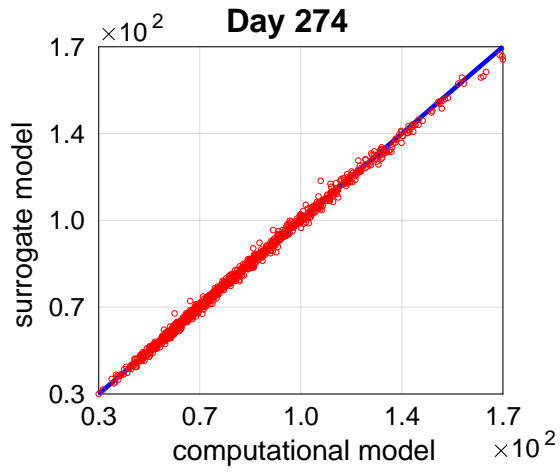


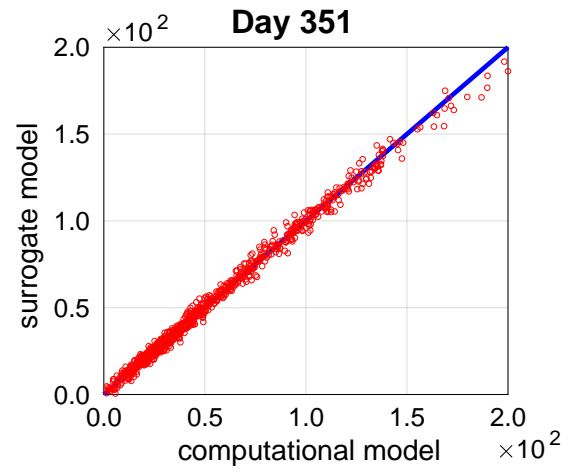
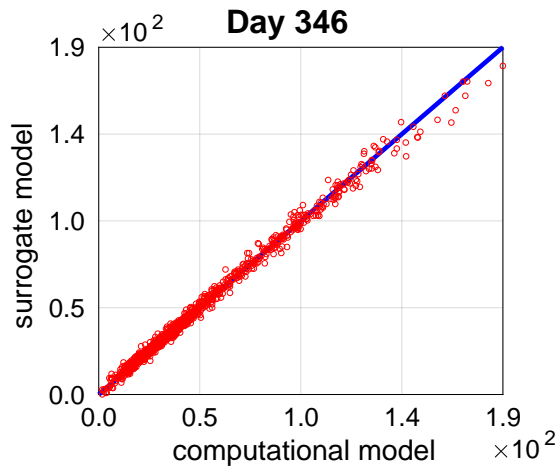
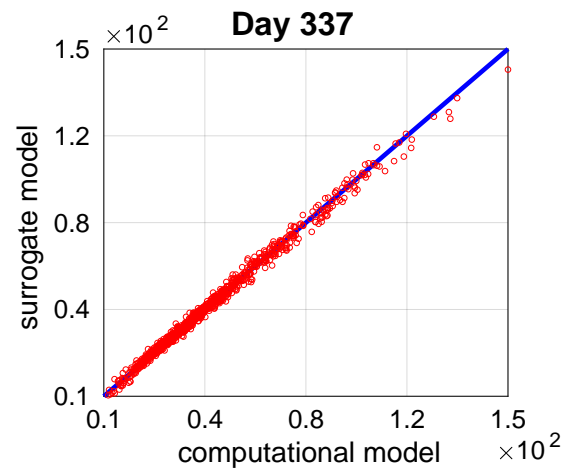
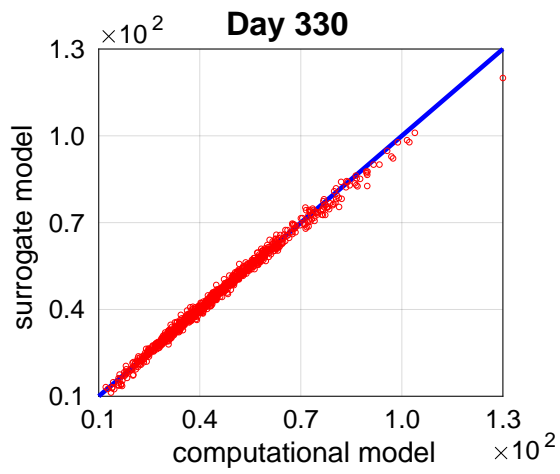
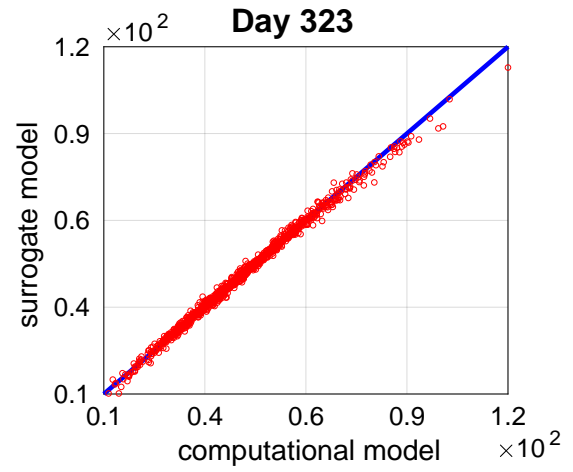
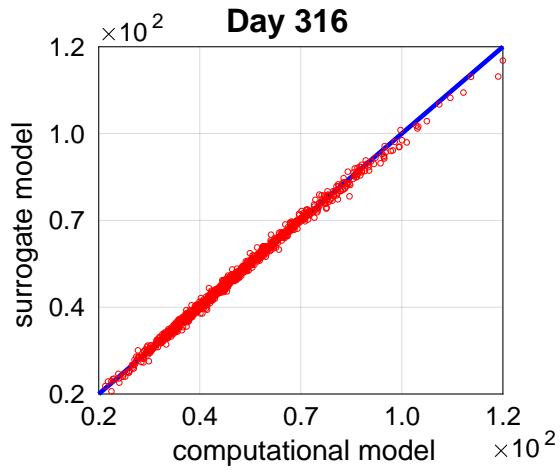


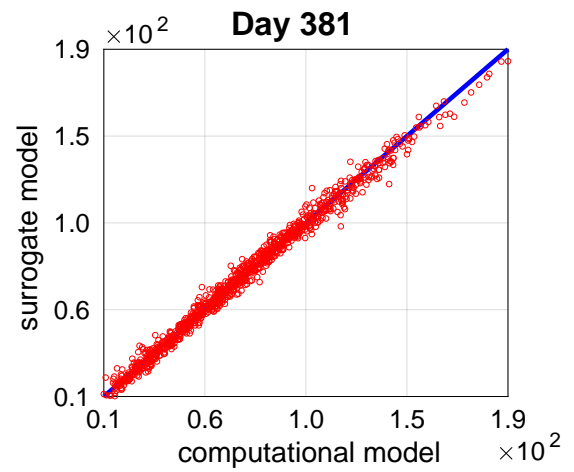
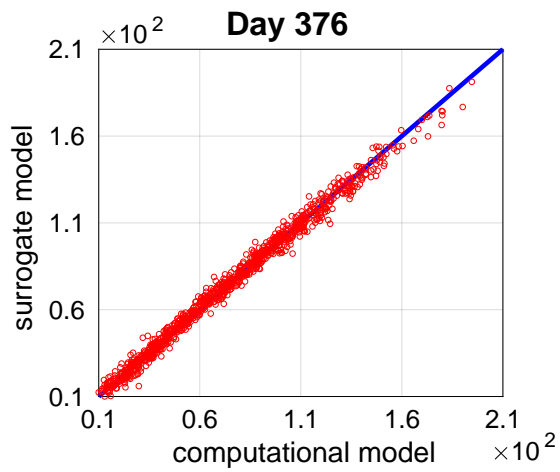
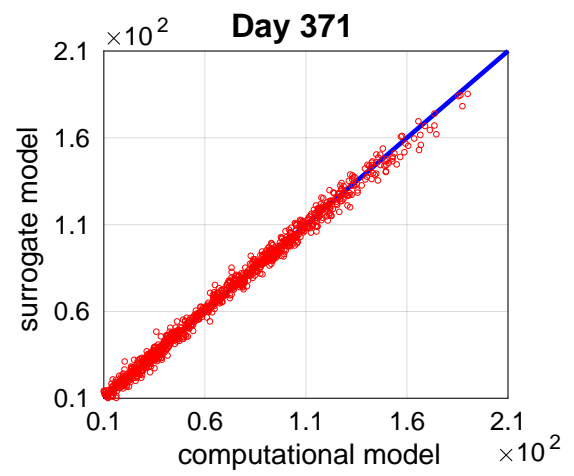
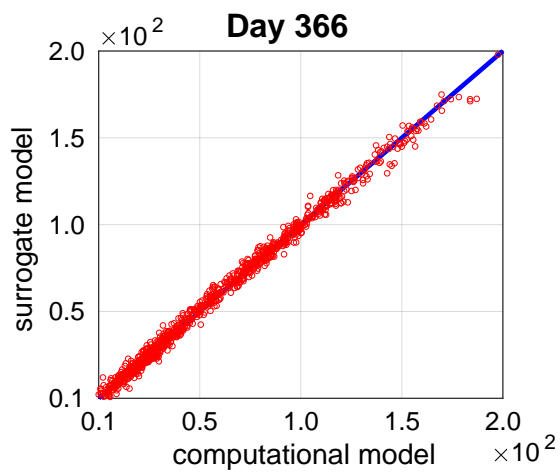
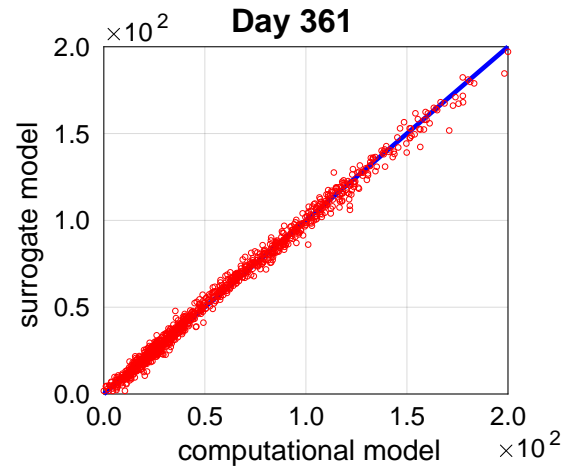
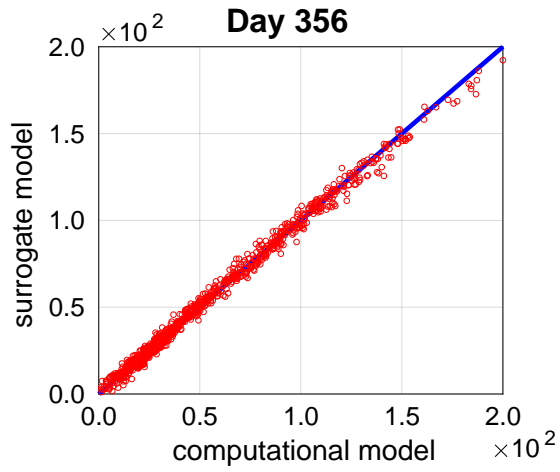


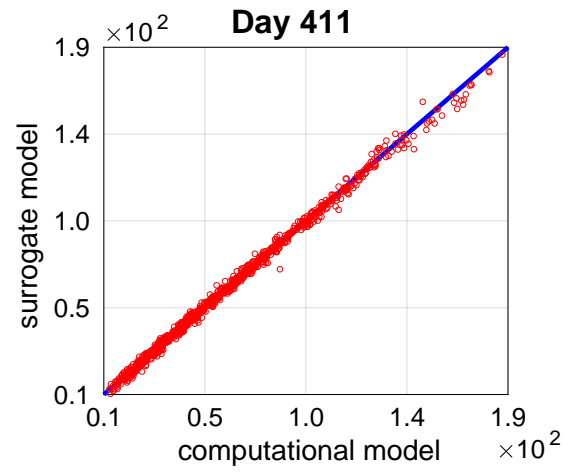
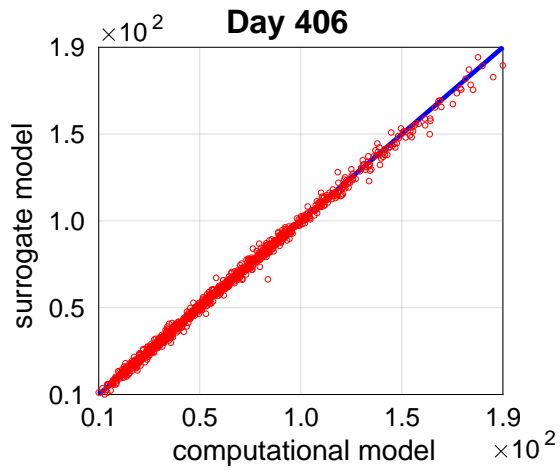
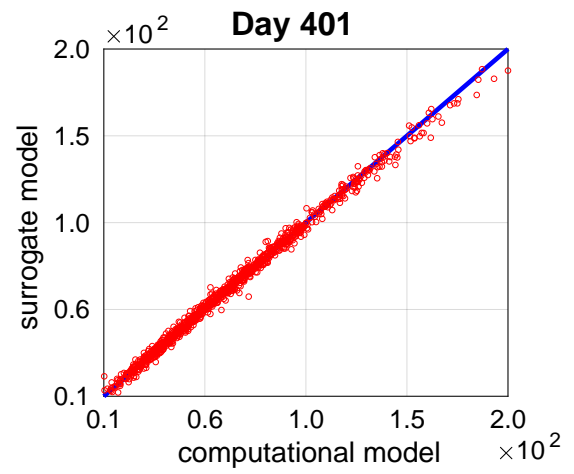
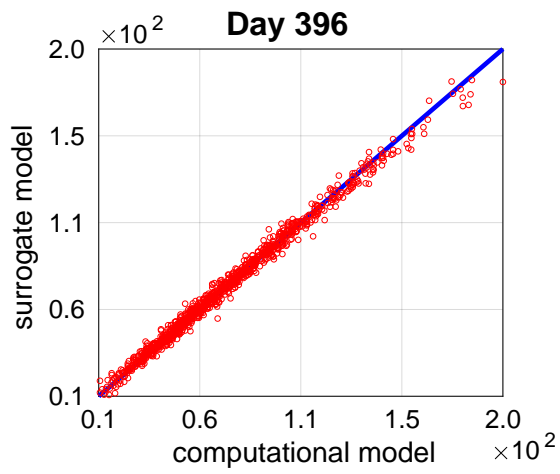
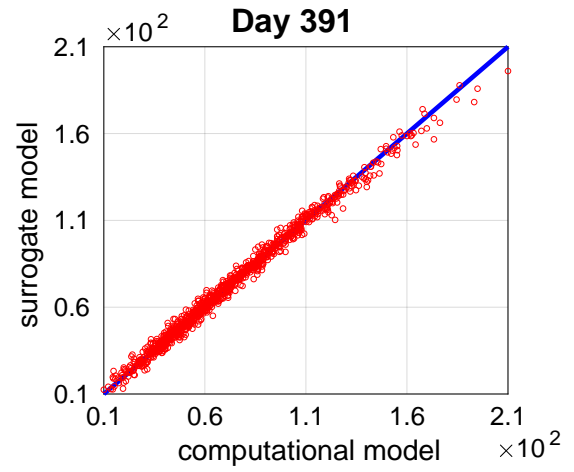
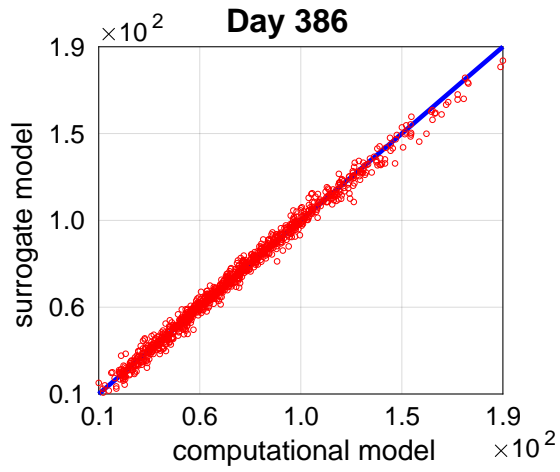


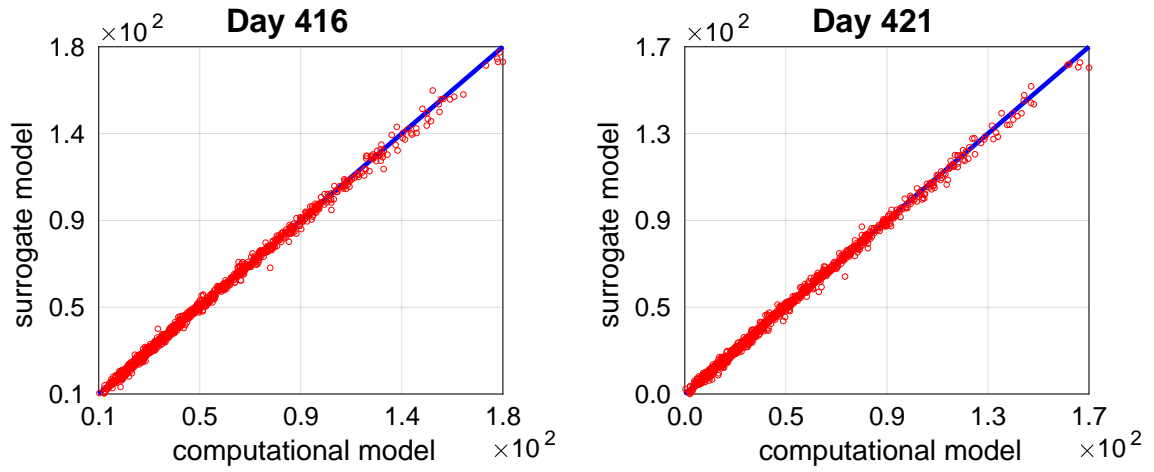




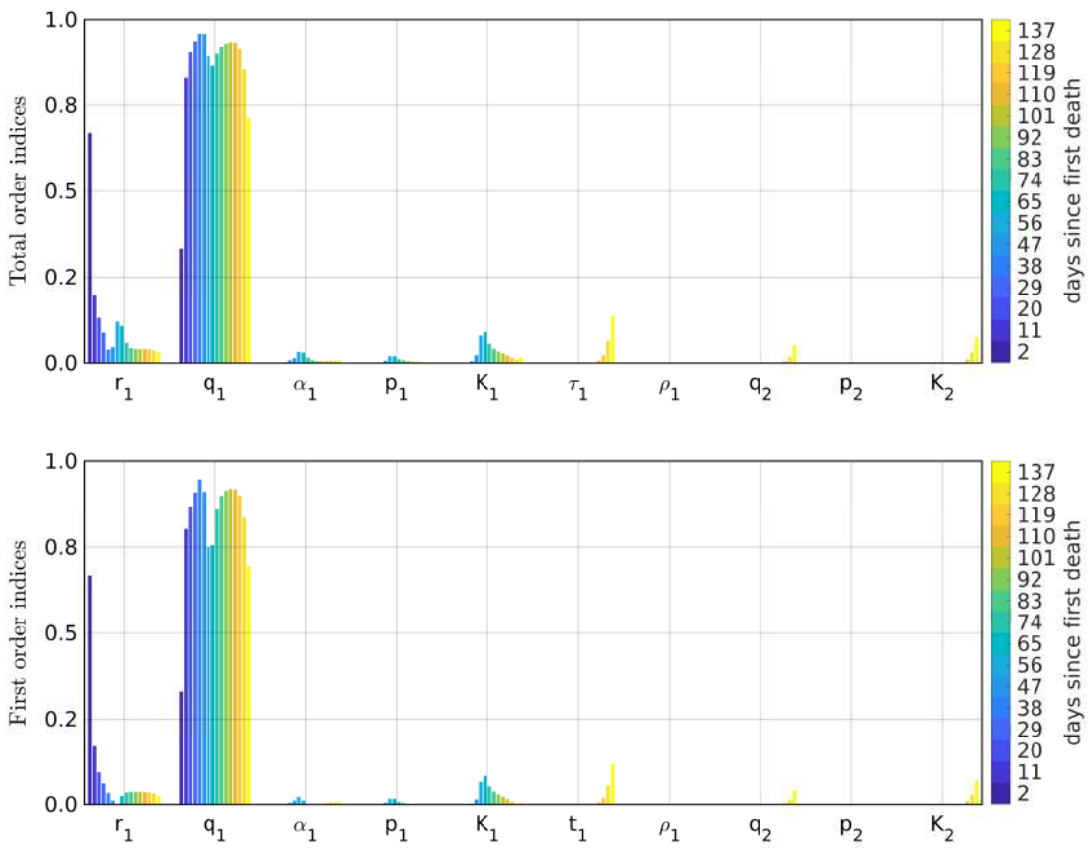


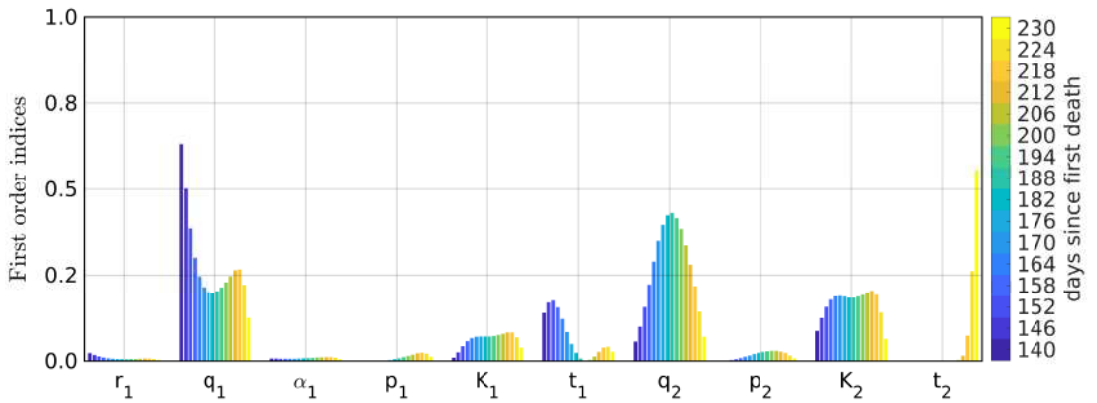
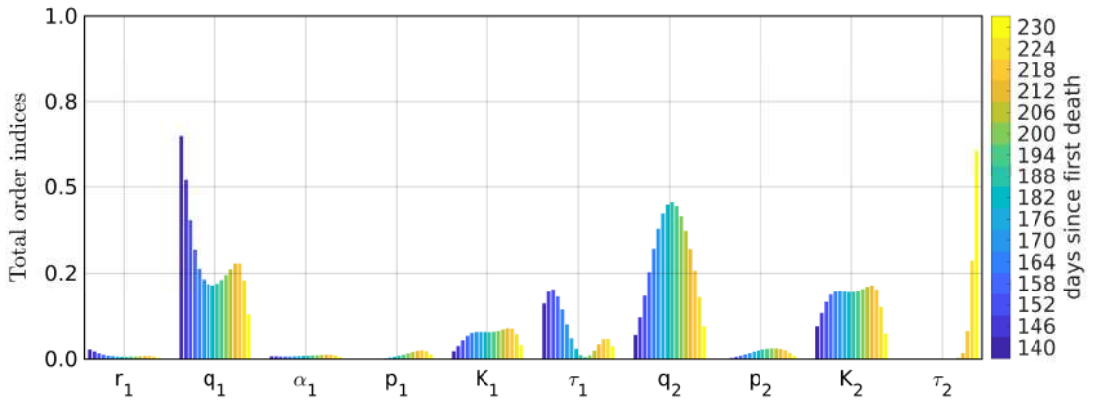
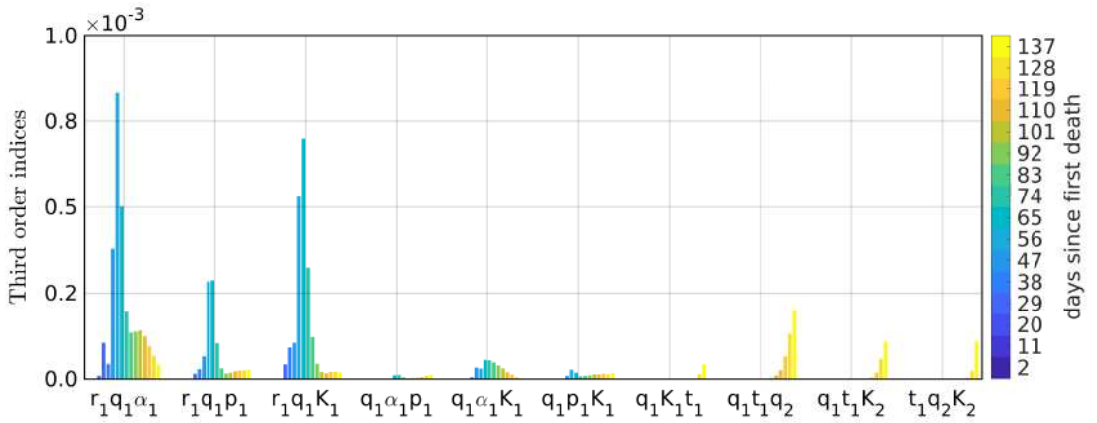
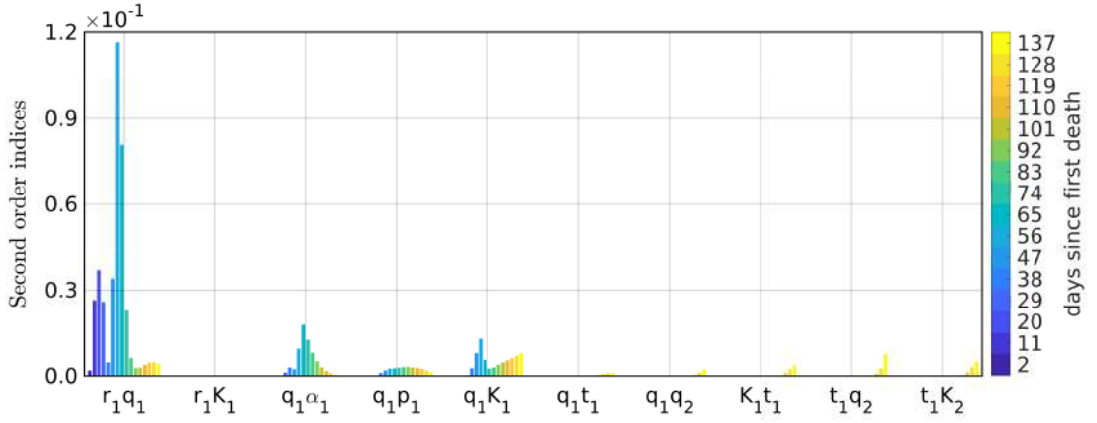


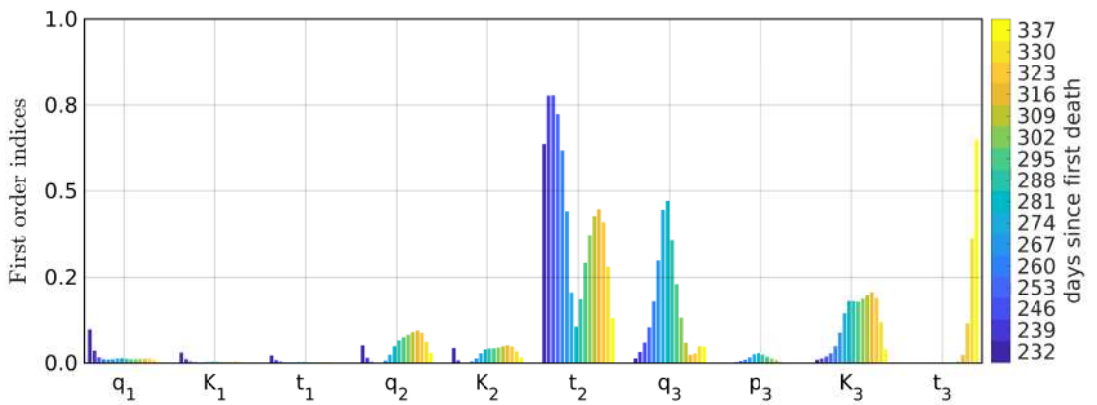
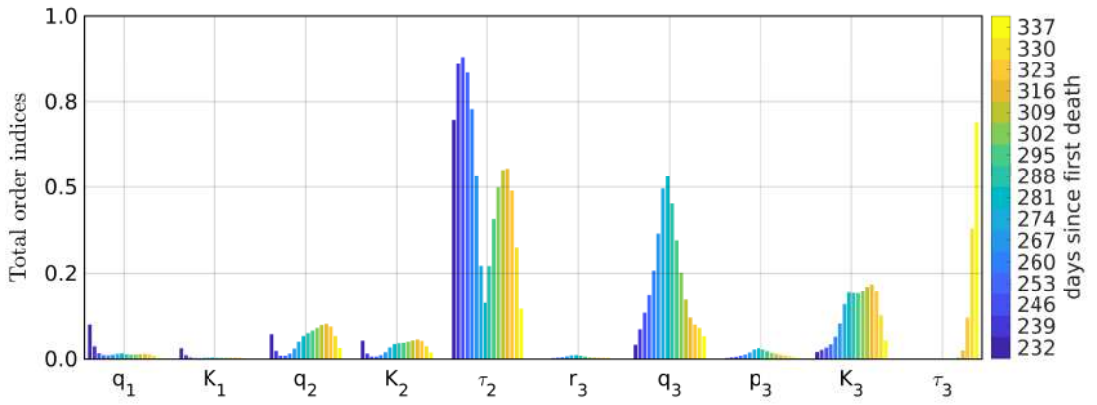
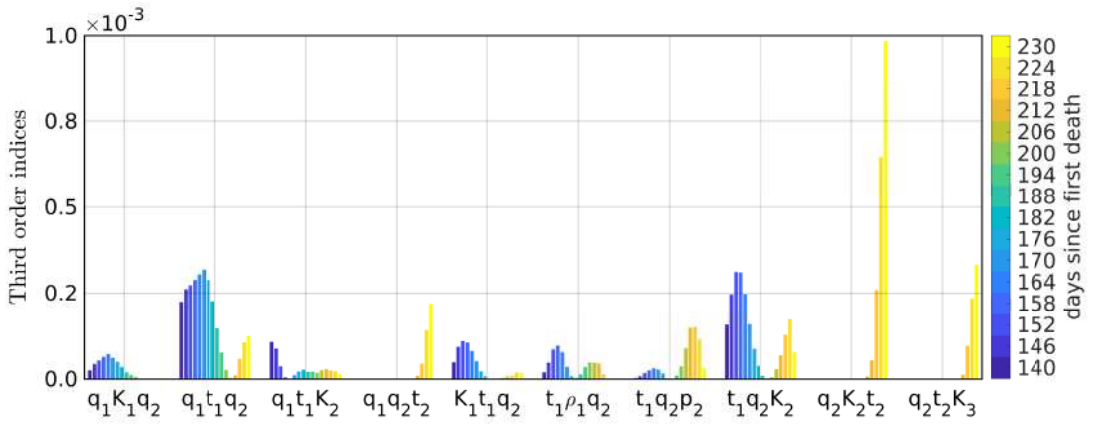
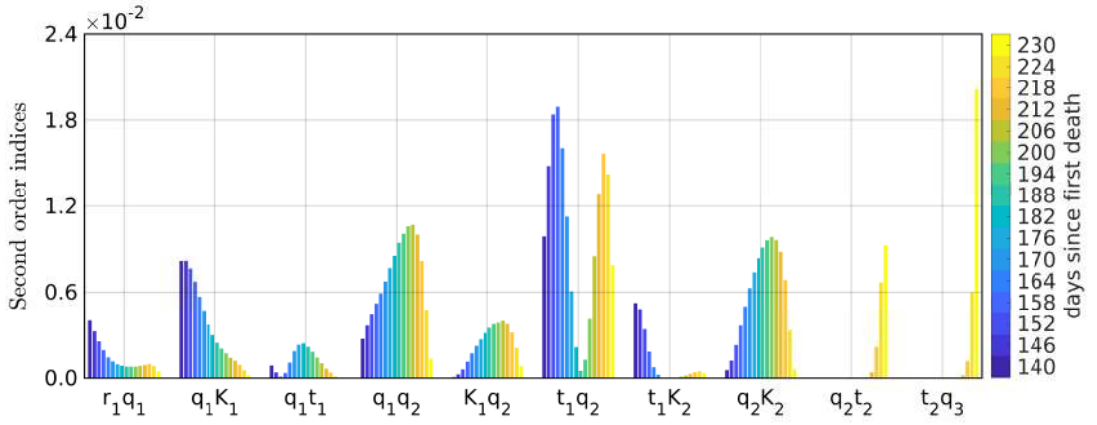


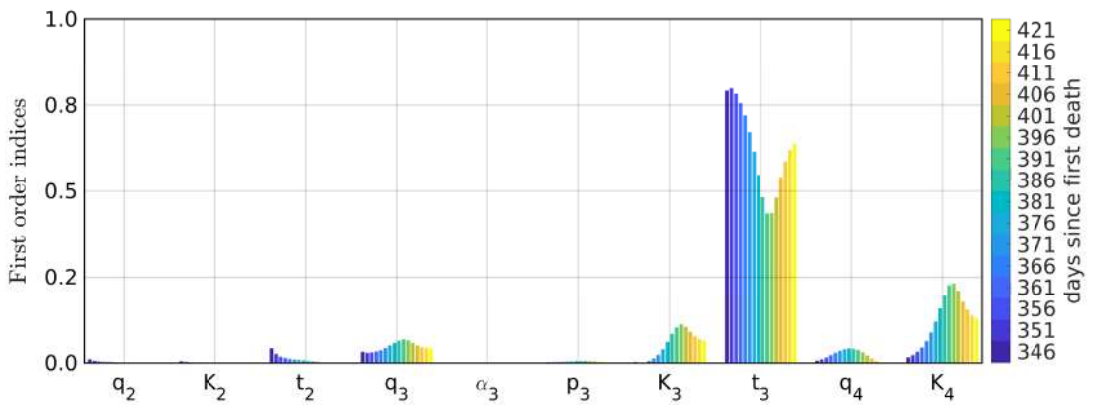
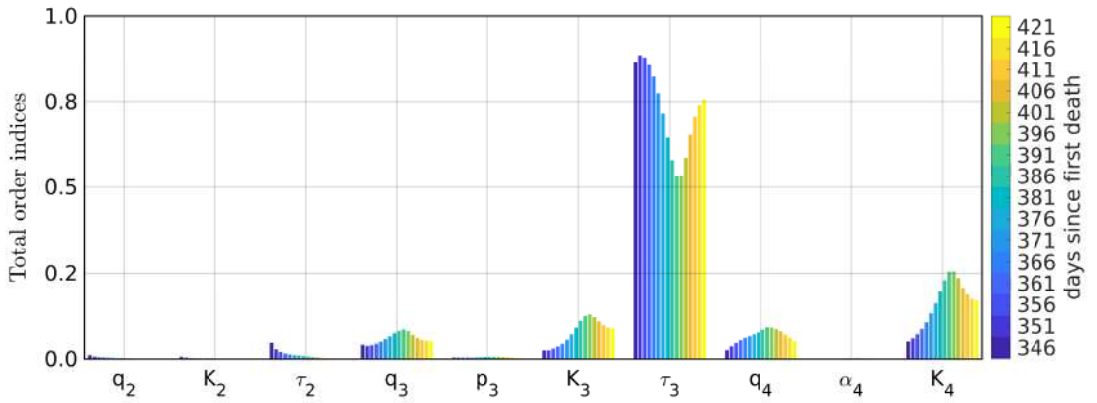
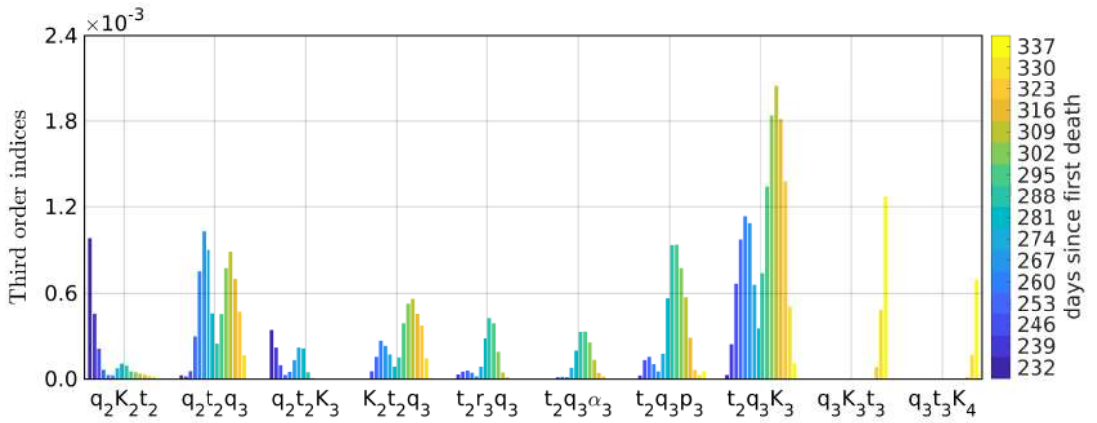
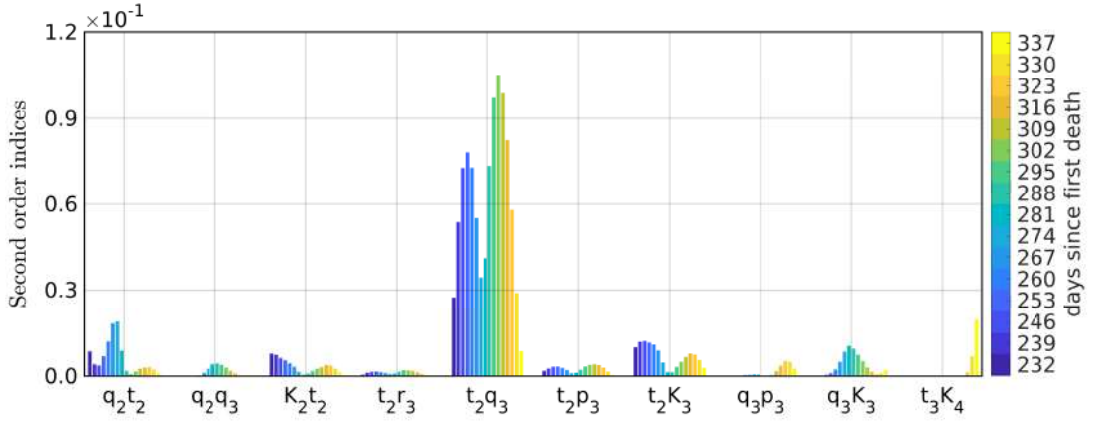


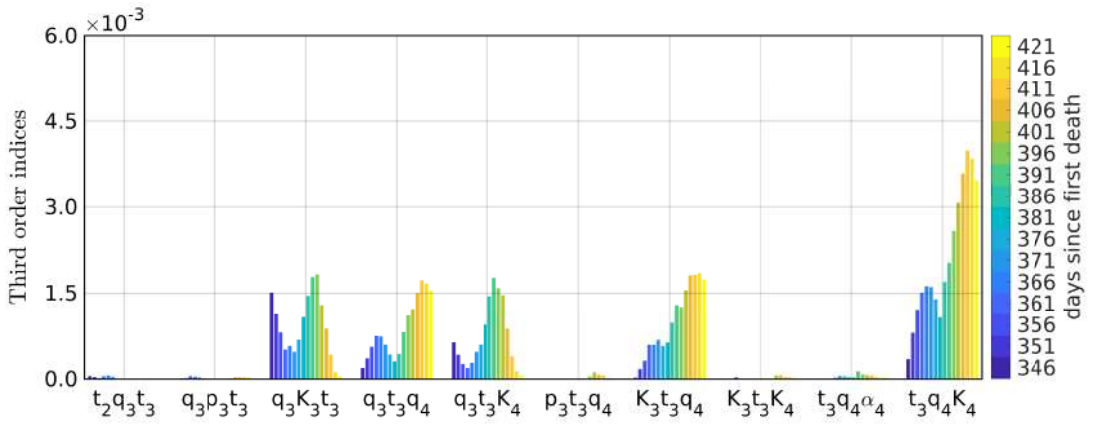
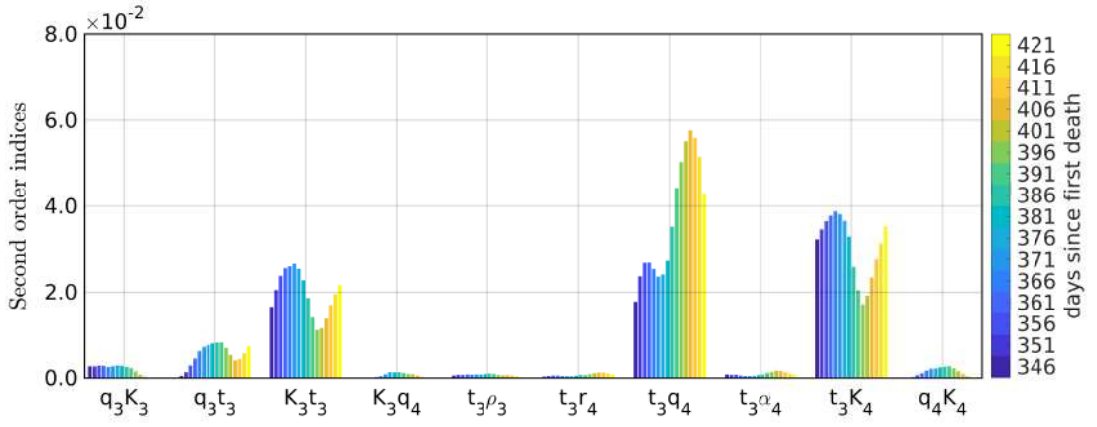
B.2.2 Sobol indices












APPENDIX C – Scientific production

C.1 Published book chapter

Chapter 6

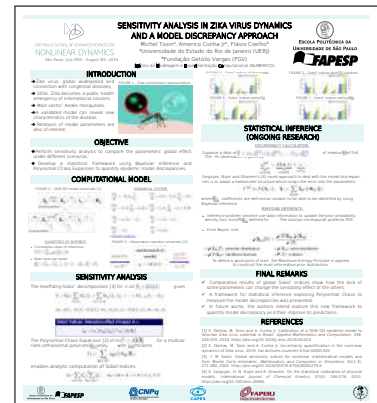
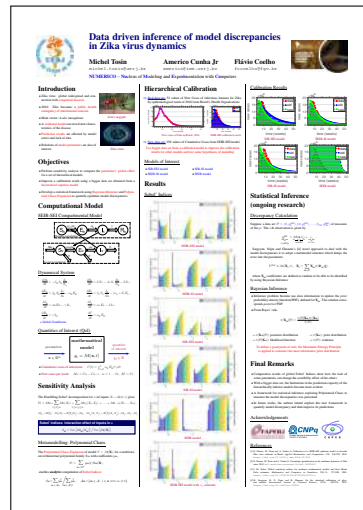
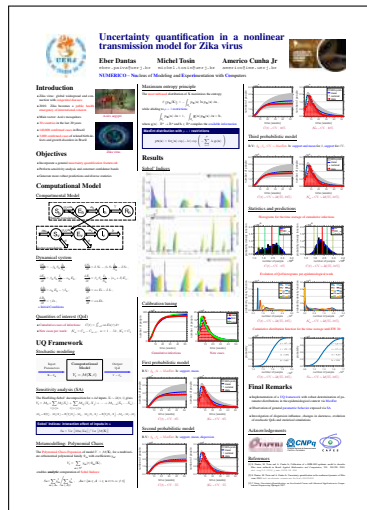
A Tutorial on Sobol' Global Sensitivity Analysis Applied to Biological Models

Michel Tosin, Adriano M. A. Côrtes, and Americo Cunha



Abstract Nowadays, in addition to traditional qualitative methods, quantitative techniques are also a standard tool to describe biological systems behavior. An example is the broad class of mathematical models, based on differential equations, used in ecology, biochemical kinetics, epidemiology, gene regulatory networks, etc. Independent of their simplicity or complexity, all these models have in common (generally unknown a priori) parameters that need to be identified from observations (data) of the real system, usually available on the literature, obtained by specific assays or surveyed by public health offices. Before using this data to calibrate the models, a good practice is to judge the most influential parameters. That can be done with aid of the Sobol' indices, a variance-based statistical technique for global sensitivity analysis, which measures the individual importance of each parameter, as well as their joint-effect, on the model output (a.k.a. quantity of interest). These variance-based indexes may be computed using Monte Carlo simulation but, depending on the model, this task can be very costly. An alternative approach for this scenario is the use of surrogate models to speed-up the calculations. Using simple biological models, from different areas, we develop a tutorial that illustrates how practitioners can use Sobol' indices to quantify, in a probabilistic manner, the relevance of the parameters of their models. This tutorial describes a very robust framework to compute Sobol' indices employing a polynomial chaos surrogate model constructed with the UQLab package.

C.2 Posters presented in events



C.3 Work presented in CBA

Seleção de modelos epidemiológicos via análise de sensibilidade global

Michel Tosin* Americo Cunha Jr** Flávio C. Coelho***

* Instituto de Matemática e Estatística, Universidade do Estado do Rio de Janeiro, RJ (e-mail: michel.tosin@uerj.br)

** Instituto de Matemática e Estatística, Universidade do Estado do Rio de Janeiro, RJ (e-mail: amercico.cunha@uerj.br)

*** Escola de Matemática Aplicada, Fundação Getúlio Vargas, RJ (e-mail: fcoelho@fgv.br)

Abstract: This paper propose a methodology for epidemiological model selection by using Akaike information criteria bringing as novelty the construction of a likelihood function based in the results of a global sensitivity analysis through the Sobol's indices obtained by using polynomial chaos expansion. The main idea is to incorporate of the information about the influence of the parameters on the response to select a more interesting model inside of a set of candidates. The strategy is applied to a set of compartmental models compatible with those used to analyze the recent COVID-19 pandemic, allowing to compare them without the presence of experimental data.

Resumo: Este paper propõe uma metodologia de seleção de modelos epidemiológicos via critério da informação de Akaike trazendo como novidade a construção de uma função de verossimilhança baseada nos resultados de uma análise de sensibilidade global através dos índices de Sobol obtidos usando expansão em polinômios caos. A ideia geral é incorporar a informação sobre a influência dos parâmetros na resposta para selecionar um modelo mais interessante dentro do conjunto de candidatos. A estratégia é aplicada a um conjunto de modelos compartimentais compatíveis aos usados para analisar a pandemia de COVID-19 recente, permitindo compará-los sem a presença de dados experimentais.

Keywords: compartmental models; COVID-19; Sobol's indices; polynomial chaos expansion; Akaike information criteria.

Palavras-chaves: modelos compartimentais; COVID-19; índices de Sobol; expansão em polinômio caos; critério da informação de Akaike.