On the merits of sparse surrogates for global sensitivity analysis of multi-scale nonlinear problems: application to turbulence and fire-spotting model in wildland fire simulators

A. Trucchia[a], V. Egorova[a], G. Pagnini[a,b], M.C. Rochoux[c]

[a] *BCAM – Basque Center for Applied Mathematics, Alameda de Mazarredo 14, 48009 Bilbao, Basque Country – Spain*
[b] *Ikerbasque – Basque Foundation for Science, Calle de María Díaz de Haro 3, 48013 Bilbao, Basque Country – Spain*
[c] *CECI, Université de Toulouse, CNRS, CERFACS, 42 Avenue Gaspard Coriolis, 31057 Toulouse cedex 1, France*

Abstract

Many nonlinear phenomena, whose numerical simulation is not straightforward, depend on a set of parameters in a way which is not easy to predict beforehand. Wildland fires in presence of strong winds fall into this category, also due to the occurrence of firespotting. We present a global sensitivity analysis of a new sub-model for turbulence and fire-spotting included in a wildfire spread model based on a stochastic representation of the fireline. To limit the number of model evaluations, fast surrogate models based on generalized Polynomial Chaos (gPC) and Gaussian Process are used to identify the key parameters affecting topology and size of burnt area. This study investigates the application of these surrogates to compute Sobol' sensitivity indices in an idealized test case. The performances of the surrogates for varying size and type of training sets as well as for varying parameterization and choice of algorithms have been compared. In particular, different types

Preprint submitted to Communications in Nonlinear Science and Numerical Simulation February 6, 2019

of truncation and projection strategies are tested for gPC surrogates. The best performance was achieved using a gPC strategy based on a sparse least-angle regression (LAR) and a low-discrepancy Halton's sequence. Still, the LAR-based gPC surrogate tends to filter out the information coming from parameters with large length-scale, which is not the case of the cleaning-based gPC surrogate. The wind is known to drive the fire propagation. The results show that it is a more general leading factor that governs the generation of secondary fires. Using a sparse surrogate is thus a promising strategy to analyze new models and its dependency on input parameters in wildfire applications.

Keywords: Sensitivity Analysis, generalized Polynomial Chaos, Gaussian Process, Wildland fire

Nomenclature

Table 1: List of abbreviations

Abbreviation	**Meaning**
ABL	Atmospheric Boundary Layer
FT	Free Atmosphere
GP	Gaussian Process
gPC	generalized Polynomial Chaos
LAR	Least Angle Regression
LSM	Level Set Method
MSR	Minimum Spanning Rectangle
PDF	Probability Density Function
ROS	Rate of Spread
SLS	Standard Least Squares
STD	STandard Deviation

Table 2: List of important static and dynamic model parameters.

Model quantities	**Units**
$\mathcal{B}(t)$, burnt area at time t	–
f, PDF of the random process	m^{-2}
$G(\mathrm{x};t)$, isotropic bivariate Gaussian PDF of turbulence	m^{-2}
$q(l)$, lognormal PDF of firebrand landing distance	m^{-1}
$\mathrm{x} = (x_1, x_2)$, horizontal space variable	m
n_{fr}, normal direction to the fireline	–
n_{U}, unit vector aligned with the mean wind direction	–
t, time	s
ϕ, level-set function	–
Ω, 2–D computational domain	–
$\lvert\Omega\rvert$, area of the computational domain	m^2
Physical Model Parameters	**Value/Units**
C_d, drag coefficient	–
D, turbulent diffusion coefficient	$\mathrm{m}^2\,\mathrm{s}^{-1}$
g, acceleration due to gravity	$9.8\ \mathrm{m\,s}^{-2}$
h, dimension of convective cell	100 m
H, fire plume height	m
I, fireline intensity	$\mathrm{kW\,m}^{-1}$
P_{f0}, reference fire power	10^6 W
U, horizontal wind vector field at mid-flame height	$\mathrm{m\,s}^{-1}$
$\lVert\mathrm{U}\rVert$, horizontal wind magnitude	$\mathrm{m\,s}^{-1}$
$\mathcal{V}$, rate of spread	$\mathrm{m\,s}^{-1}$
z_p, pth percentile	0.45
Δh_c, heat of combustion of wildland fuels	18,620 $\mathrm{kJ\,kg}^{-1}$
(μ, σ) , parameters of the log-normal PDF $q(l)$	–
ρ_a, air density	1.2 $\mathrm{kg\,m}^{-3}$
ρ_f^*, wildland fuel density (*Pinus Ponderosa*)	542 $\mathrm{kg\,m}^{-3}$
ω_0, oven-dry mass of wildland fuel	2.243 $\mathrm{kg\,m}^{-2}$
τ, ignition delay of firebrands	s
χ, air thermal diffusivity	$2 \cdot 10^{-5}\,\mathrm{m}^2\,\mathrm{s}^{-1}$
ΔT, temperature difference of convective cell	800-923 K
ℓ, firebrand landing distance	m
ν, kinematic viscosity	$1.5 \cdot 10^{-5}\,\mathrm{m}^2\,\mathrm{s}^{-1}$
γ, thermal expansion coefficient	K^{-1}
$\alpha_H, \beta_H, \gamma_H, \delta_H$, coefficients for fire plume height H	–

Table 3: List of important algorithmic parameters.

A_t, burnt area ratio at time t
d, dimension of the stochastic space ($d = 3$)
$\mathcal{D}_N$, training set of size N
$\mathcal{M}$, forward model
$\mathcal{M}_{\mathrm{pc}}$, gPC-expansion
N, size of the training set
P, total polynomial order
q, hyperbolic truncation parameter
r, number of terms in the surrogate basis
S_t, minimum spanning rectangle ratio at time t
y, quantity of interest
$\widehat{\mathrm{y}}$, estimate of the quantity of interest y
$\mathrm{y}^{(k)}$, kth realization of the quantity of interest y
$\mathcal{A}$, set of selected multi-indices in gPC-expansion
$\boldsymbol{\alpha}$, multi-index for gPC-expansion
δ, Kronecker delta-function
$\boldsymbol{\theta} = (\theta_1, \cdots, \theta_d)$, vector of uncertain input parameters, $[\|\mathrm{U}\|, I, \tau]$ or $[\mu, \sigma, D]$
$\boldsymbol{\theta}^{(k)}$, kth realization of the uncertain input vector $\boldsymbol{\theta}$
$\boldsymbol{\zeta} = (\zeta_1, \cdots, \zeta_d)$, vector $\boldsymbol{\theta}$ in standard probabilistic space
ρ_{θ_i}, marginal PDF of ith input parameter in $\boldsymbol{\theta}$
$\boldsymbol{\rho}_{\boldsymbol{\zeta}}$, joint PDF of $\boldsymbol{\theta}$ in standard probabilistic space
$\Psi_{\boldsymbol{\alpha}}$, $\boldsymbol{\alpha}$th basis function for surrogate model
Φ_{α_i}, ith one-dimensional basis function
$\gamma_{\boldsymbol{\alpha}}$, $\boldsymbol{\alpha}$th coefficient in the surrogate basis
$\boldsymbol{\gamma}$, vector of surrogate coefficients
$\left(\omega^{(k)}, \boldsymbol{\zeta}^{(k)}\right)$, kth quadrature weight and root
ℓ_{gp}, correlation length-scale for GP-model
σ_{gp}, observable standard deviation for GP-model
τ_{gp}, nugget effect for GP-model
$\pi(\boldsymbol{\theta}, \boldsymbol{\theta}')$, correlation kernel for GP-model
ϵ_{emp}, empirical training error
Q_2, cross-validation predictive coefficient

1. Introduction

Despite our recent progress in computer-based wildland fire spread modeling and remote sensing technology, our general understanding of wildland fire behavior remains limited. This is mainly due to the complexity of wildfire dynamics that results from multi-scale interactions between biomass pyrolysis, combustion and turbulent flow dynamics, heat transfer as well as atmospheric dynamics [1, 2, 3, 4, 5, 6]. Turbulence plays an important role: wildland fires release large amounts of heat that lead to the development of a turbulent flow in the vicinity of the flame zone and thereby enhance the heat transfer to unburnt fuel, boosting biomass fuel ignition, combustion and fire spread. There is therefore a strong coupling between wildland fires and micrometeorology [7, 8, 9, 10, 11, 12]. When extreme conditions are met in complex terrain such as canyons in combination with strong winds and severe drought, highly destructive fires referred to as "megafires" can develop [13, 14, 15, 16]. For such fires, a massive buoyant smoke plume forms above the flame zone modifying micro-meteorological conditions [17] and thereby fire spread conditions. Windborne embers can be transported over large distances, causing fire spotting and further ignitions downstream from the current fire, leading to multiple "spot fires" that are difficult to stop by firefighters and that dramatically increase fire danger. Turbulence and fire-spotting result from very nonlinear effects that are still poorly understood and that have been identified as a valuable research target with direct applications in fire emergency response, especially at wildland-urban interface [18].

The representation of these processes is beyond the scope of current operational wildfire spread models. At regional scales (i.e. at scales ranging from a few tens of meters up to several hectares), a wildland fire is indeed represented as a two-dimensional propagating interface (referred to as the "fire front" or "fireline") separating the burnt area to the unburnt vegetation; the local propagation speed is called the "rate of spread" (ROS). This front representation is the dominant approach in current wildfire spread simulators such as FARSITE [19], FOREFIRE [20, 12], PROMETHEUS [21], PHOENIX RapidFire [22], SFIRE [11] or ELMFIRE [23]. These simulators rely on an empirical parametrization of the ROS that is derived from steady-state assumption and that is an analytic function of biomass fuel properties, topographical properties and micro-meteorological conditions [24]. The ROS submodel is included in an Eulerian or Lagrangian front-tracking solver to simulate the fireline propagation. This approach is limited in scope [25, 26, 27] due to the large uncertainties associated with the input parameters of the ROS model [28, 29], which can be partially reduced by integrating real-time fire front measurements through data assimilation [11, 30, 31, 32, 33, 34, 35, 36]. This approach is also limited due to the lack of knowledge on the physics of the fire problem [5], in particular on the processes associated with turbulence and fire-spotting.

These modeling limitations at regional scales have motivated investigation of turbulence and fire-spotting effects both from experimental and modeling viewpoints [37, 38, 39, 40, 18, 41, 42, 43, 44]. To better characterize these nonlinear processes, there is a need to develop new submodels including the effects of random processes such as turbulence and fire-spotting in operationally-oriented wildfire spread models. This is one of the objectives of the work proposed in [45, 46, 47, 48], which introduces a randomized representation of the fireline. A novel family of reaction-diffusion equations have been developed to link front models to reaction-diffusion ones and thereby integrate the effects of random processes in fire models. The front propagation is randomized by adding to the driving function, a random displacement distributed according to a probability density function (PDF) corresponding to heat turbulent transport and fire-spotting landing distance. The driving equation of the resulting averaged process is analogous to an evolution equation of the reaction-diffusion type, where the ROS controls the source term. In absence of random processes, the model is identical to the one given by the standard wildfire spread model, which is only driven by the ROS analytic function.

Including new modeling components in wildfire spread simulators adds some complexity and in particular introduces new model parameters. There is therefore a strong need to perform sensitivity analysis to analyze in a rigorous way the model structure, i.e. the dependency between the input parameters and the simulated quantities of interest (here, the topology and the extension of the burnt area at a given time). The objective in such an extensive global sensitivity analysis is two-fold. First, sensitivity analysis identifies the most influential parameters on the model predictions over a wide range of values for the model parameters, ranks them by order of importance and spots unimportant parameters [49, 50, 51]. This is helpful to provide hints and guidelines about the physical processes that are essential to account for to track wildland fire behavior. Second, sensitivity analysis is a mandatory step to select which are the estimation targets to consider when the wildfire spread model is integrated in a data assimilation framework to produce short-term predictions of wildfire behavior; the model parameters shall indeed be uncertain and the quantities of interest shall be sensitive to changes in these model parameters to ensure data assimilation is efficient [52, 32, 35].

When relying on stochastic non-intrusive methods (meaning that no modification of the physical model, also referred to as the "forward model", is required), global sensitivity analysis requires the use of an ensemble of model evaluations. This procedure can be divided into three steps: *i)* characterization of the variability in the model parameters based on available information and statistical sampling to obtain an ensemble of parameter values; *ii)* multiple evaluations of the forward model while accounting for the identified uncertainties to obtain an ensemble of quantities of interest (the forward model is used as a "black-box"); and *iii)* computing Sobol' sensitivity indices [53] that provides a relative measure of how the variability of the model response is affected by the variability in each uncertain parameter (this variability is measured in terms of variance). Computing these Sobol' indices therefore requires to have access to an accurate mapping between the uncertain inputs and the quantities of interest. This is computationally intensive when using standard Monte Carlo sampling method since this method features a slow convergence rate and thus requires a large ensemble to obtain reliable statistics. The cost of global sensitivity analysis is significantly reduced when the forward model is replaced by a surrogate model that mimics its response for the considered range of the model parameters. The formulation of such a surrogate requires a limited number of model evaluations, referred to as the "training set". Then the surrogate can be evaluated multiples times at almost no cost to evaluate uncertainties in the quantities of interest and/or perform sensitivity analysis [54, 55, 56, 57, 58].

There are various ways of formulating a surrogate. In the present work, we focus our attention on generalized polynomial chaos (gPC) expansions [59, 60, 61, 54, 62] and Gaussian process (GP) models [63, 64, 57, 55, 65, 66, 67]. The gPC-approach formulates a polynomial expansion, in which the basis is defined according to the PDF of the uncertain parameters and in which the associated weights directly relate to the statistics of the quantities of interest. This implies that by construction the quantities of interest are projected upon the same basis as the input parameters. The GP-approach adopts a different viewpoint by considering the simulated quantities of interest as a realization of a Gaussian stochastic process conditioned by the training set. This stochastic process is fully characterized with mean and covariance kernel functions, which rely on the estimation of hyperparameters. Both gPC and GP surrogates are compared in the literature for uncertainty quantification and sensitivity analysis studies [57, 58, 68, 69]. Still, the ranking between gPC and GP approaches remains problem-dependent. It is thus of great interest to compare these approaches for application in wildland fires.

In wildland fire applications, the performance of the gPC-approach has already been demonstrated within the framework of data assimilation to reduce the computational cost of sequential parameter estimation [32, 36]. However, the gPC-algorithm relied on the use of a full basis and a standard spectral projection method. Building the surrogate this way may be too costly for high-dimensional problems, i.e. when the number of uncertain parameters increases. There exists more advanced gPC-strategies in the literature to reduce the number of elements in the gPC basis and thus reduce the required size of the training set, see [70, 71, 72]. Due to the multiple sources of uncertainty in wildland fire models, there is a strong need to evaluate the performance of gPC and GP approaches, i.e. for varying size and type of the training set as well as for varying parameterization and choice of the surrogate algorithms. In the present study, the objective is to determine what is the best surrogate strategy to compute Sobol' sensitivity indices and thereby examine the relevance of the parameters that are part of the turbulence and fire-spotting submodel included in the wildfire spread model [47]. Our objective is to identify the key parameters affecting the topology and the size of the burnt area that is simulated by an Eulerian-type fire spread model (`LSFire+`) and that corresponds to an idealized test case. For this purpose, we compare the performance of gPC-expansion and GP-model in their standard and sparse versions for a fixed size of the training set with different designs of experiment (Monte Carlo random sampling, quasi-random Halton's sequence, quadrature rule); a convergence study is carried out to determine the required size of the training set to ensure accuracy.

The structure of the paper is as follows. Section 2 introduces the wildfire spread model, the main sources of uncertainty, the quantities of interest and the idealized test case study. The gPC and GP approaches are detailed in Section 3 along with statistical analysis tools and error metrics. Section 4 presents the results of the comparative study between gPC and GP algorithms for different types of truncation, projection and training set. Conclusions and perspectives are given in Section 5.

2. Wildland Fire Model and Sources of Uncertainties

2.1. Forward Model

We focus the present study on Eulerian-type wildfire spread model (`LSFire+`) based on level-set methods [73, 74, 75]. This is similar to the approach adopted in the ELMFIRE fire simulator [23, 76] or the WRF-SFIRE coupled fire-atmosphere system [11].

2.1.1. Deterministic Front Propagation

To represent the time-evolving burning active areas over the computational domain $\Omega \subset \mathbb{R}^2$, we introduce an implicit function $\phi \equiv \phi(\mathrm{x}, t)$ as the fireline marker with $\phi : \Omega \times [0; +\infty[\to \mathbb{R}$. The fireline is identified as the contour line $\phi(\mathrm{x}, t) = \phi^*$ referred to as the "level set". We thus denote the time-evolving two-dimensional burnt area as $\mathcal{B}(t) = \{\mathrm{x} = (x_1, x_2) \in \Omega \,|\, \phi(\mathrm{x}, t) > \phi^*\}$.

The temporal evolution of the level set $\phi(\mathrm{x}, t) = \phi^*$ is governed by the Eikonal equation

$$\frac{\partial \phi}{\partial t}(\mathrm{x}, t) = \mathcal{V}(\mathrm{x}, t)\, \|\nabla \phi(\mathrm{x}, t)\|\,, \quad \phi(\mathrm{x}, 0) = \phi_0(\mathrm{x}), \quad \mathrm{x} \in \Omega,\ t \geq 0, \qquad (1)$$

where $\mathcal{V}$ corresponds to the ROS parameterization that is a function of the wind field $\mathrm{U}(\mathrm{x}, t)$, orography and biomass fuel conditions, and where $\phi_0(\mathrm{x})$ is the initial condition at time 0. The propagation of the fireline is assumed to be directed towards the normal direction to the front.

2.1.2. Random Front Formulation

The stochastic approach that is adopted in the present study is based on the idea of splitting the motion of the fireline into a drifting part and a fluctuating part [47, 77, 48]. The drifting part corresponds to the resolution of the deterministic problem in Eq. (1). The fluctuating part results from a comprehensive statistical description of the dynamic system, which includes random effects in agreement with the physics of the system.

The motion of each burning point can be random due to the effect of turbulence and/or fire-spotting. The effective indicator function, $\phi_{\mathrm{e}}(\mathrm{x}, t) : \mathcal{B} \times [0, +\infty[\rightarrow [0, 1]$ emerges from the superposition of the front weighted by the distribution of fluctuations around the deterministic front, i.e.

$$\phi_{\mathrm{e}}(\mathrm{x}, t) = \int_{\mathcal{B}} \phi(\overline{\mathrm{x}}, t)\, f(\mathrm{x}; t|\overline{\mathrm{x}})\, d\overline{\mathrm{x}}, \tag{2}$$

where $f(\mathrm{x}; t|\overline{\mathrm{x}})$ denotes the PDF of the displacement of the active burning points around the mean position $\overline{\mathrm{x}}$. An arbitrary threshold value $\phi_{\mathrm{e,fr}}$ is used as the criterion to separate burnt area and unburnt area. The effective burnt area is therefore defined as $\mathcal{B}_{\mathrm{e}}(\mathrm{x}, t) = \{\mathrm{x} \in \mathcal{B} \mid \phi_{\mathrm{e}}(\mathrm{x}, t) > \phi_{\mathrm{e,fr}}\}$.

Note that the PDF $f(\mathrm{x}; t|\overline{\mathrm{x}})$ is associated with two independent random variables representing turbulence and fire-spotting, with fire-spotting a downwind phenomenon acting along the wind direction. $f(\mathrm{x}; t|\overline{\mathrm{x}})$ is expressed as

$$f(\mathrm{x}; t|\overline{\mathrm{x}}) = \begin{cases} \displaystyle\int_0^{\infty} G(\mathrm{x} - \overline{\mathrm{x}} - \ell\, \mathrm{n}_U; t)\, q(\ell; t)\, d\ell\,, & \mathrm{n} \cdot \mathrm{n}_{\mathrm{U}} \geq 0\,, \\ \\ G(\mathrm{x} - \overline{\mathrm{x}}; t), & \text{otherwise}\,, \end{cases} \tag{3}$$

where $\mathrm{n_U}$ is the unit vector aligned with the mean wind direction, where $G(\mathrm{x}-\overline{\mathrm{x}};t)$ is the PDF associated with turbulent diffusion, and where $q(\ell;t)$ is the PDF associated with firebrand landing distance ℓ. We follow the same choices as in [47, 77, 48]. Hence, we assume that turbulent diffusion is isotropic and represented as a bivariate Gaussian PDF

$$G(\mathrm{x}-\overline{\mathrm{x}};t) = \frac{1}{4\pi\, D\, t} \exp\left\{ \frac{(x_1-\overline{x}_1)^2+(x_2-\overline{x}_2)^2}{4\, D\, t} \right\}, \tag{4}$$

where D is the turbulent diffusion coefficient. We also assume that the downwind distribution of the firebrands follows a log-normal distribution

$$q(\ell;t) = \frac{1}{\sqrt{2\pi}\,\sigma\,\ell} \exp\left\{ -\frac{(\ln \ell/\ell_0 - \mu)^2}{2\,\sigma^2} \right\}, \tag{5}$$

where $\mu \equiv \mu(t) = \langle \ln \ell/\ell_0 \rangle$ and $\sigma \equiv \sigma(t) = \sqrt{\langle (\ln \ell/\ell_0 - \mu)^2 \rangle}$ are the mean and the standard deviation (STD) of $\ln \ell/\ell_0$, respectively, and where ℓ_0 is a unit reference length.

Since fuel ignition due to hot air and firebrands is not instantaneous, a suitable criterion related to ignition delay is introduced. This criterion is based on heating-before-burning mechanism as follows:

$$\psi(\mathrm{x},t) = \int_0^t \phi_{\mathrm{e}}(\mathrm{x},\eta)\,\frac{d\eta}{\tau}, \tag{6}$$

where $\psi(\mathrm{x},0)=0$ corresponds to the initial unburnt biomass fuel, and where τ is a reference time for ignition delay. A point x is considered ignited at time t when $\psi(\mathrm{x},t)=1$. In this case, $\mathrm{x} \in \mathcal{B}(t)$.

2.1.3. Rate of Spread Submodel and Test Case Study

Since the focus is here on sensitivity analysis methodology, we consider a simplified version of the ROS parameterization required in Eq. (1). The maximum value of the ROS, $\mathcal{V}(\mathrm{x}, t)$, is specified by means of Byram's formula [78, 79]:

$$\mathcal{V}_0 = \frac{I}{\Delta h_c \, \omega_0}, \tag{7}$$

where I [$\mathrm{kW\,m^{-1}}$] is the fireline intensity, Δh_c [$\mathrm{kJ\,kg^{-1}}$] is the fuel heat of combustion and ω_0 [$\mathrm{kg\,m^{-2}}$] is the oven-dry mass of fuel consumed per unit area in the active flaming zone. By analogy to the approach adopted in [47], the effect of the near-surface wind U on the ROS is accounted for through a corrective factor f_w as follows:

$$\mathcal{V} = \mathcal{V}_0 \, \frac{(1 + f_\mathrm{w})}{\alpha_\mathrm{w}}, \tag{8}$$

where f_w is computed following the choices made in the `fire-Lib` and `Fire Behaviour SDK` libraries (`http://fire.org`; see also [11], in the case of the NFFL – Northern Forest Fire Laboratory – Model 9), and where α_w is a suitable angle parameter for ensuring that the maximum ROS in the upwind direction is equal to the ROS prescribed by Byram's formula (7).This choice makes the ROS dependent on the wind direction rather than on its magnitude to constrain the well-known dominant role of the wind in the fire propagation and to allow for the emergence, if they exist, of second-order effects due to other factors.

In the present study, we consider an idealized test case of wildland fire. The computational domain is $7,200$ m $\times$ $6,000$ m. Terrain is flat. Vegetation is homogeneous. The wind is uniform and constant. Fire ignition is represented as a circular front characterized by a radius $r_c = 130$ m and a center located at $\mathrm{x}_c = (1,500 \text{ m}; 3,000 \text{ m})$.

2.2. Model Input Description

The set of uncertain parameters is noted $\boldsymbol{\theta} \in \mathbb{R}^d$, where d is the number of parameters to consider for sensitivity analysis. We consider two different sets of uncertain model parameters in the present work with $d = 3$. To carry out sensitivity analysis, we need to prescribe a PDF representing the statistics of each parameter and thereby its variability; this corresponds to step. *i)* discussed in the Introduction.

2.2.1. Sensitivity analysis for macroscopic/microscopic quantities

The first set of parameters mixes macroscopic and microscopic quantities: the wind speed magnitude $\|U\|$, the fireline intensity I and the ignition delay τ. Sensitivity analysis with $\boldsymbol{\theta} = (\|\mathrm{U}\|, I, \tau)^T$ corresponds to a preliminary step: we consider uniform marginal distributions that spanned around the mean values adopted in previous work [47, 77, 48], see Table 4.

Table 4: Ranges of variation and uniform marginal PDFs for $\boldsymbol{\theta} = (\|\mathrm{U}\|, I, \tau)^T$. Note that the uniform distribution is formulated as $\mathcal{U}(a; b)$ with a the minimum value and b the maximum value of the parameter.

Parameter	Uniform distribution
Wind $\|\mathrm{U}\|$ [m s^{-1}]	$\mathcal{U}(6; 14)$
Fireline intensity I [kW m^{-1}]	$\mathcal{U}(15,000; 25,000)$
Reference time for ignition delay τ [s]	$\mathcal{U}(0.6; 1.4)$

2.2.2. Sensitivity analysis for microscopic parameters

The focus of the present work is to explore the dependence of the wildfire spread model on a set of microscopic variables. We therefore determine a suitable Bayesian description for the uncertain parameters $\boldsymbol{\theta} = (\mu, \sigma, D)^T$, which relate exclusively to the fluctuating part of the forward model. Recall that μ and σ are two parameters of the log-normal PDF $q(\ell; t)$ (Eq. 5) that describes the ember landing position. Recall also that D is the diffusive coefficient of turbulent hot air involved in the Gaussian PDF $G(\mathrm{x} - \overline{\mathrm{x}}; t)$ (Eq. 4) that describes turbulent diffusion. Some functional dependence is explored for each parameter and their marginal PDFs are determined using a Monte Carlo random sampling. The resulting Beta-distributions are summarized in Table 5.

Physical parameterization. We assume that all turbulent processes are represented in the forward model through the standalone turbulent diffusion coefficient D. We only consider turbulent fluctuations, implying that the estimation of D is independent of the wind U. Since we consider a flat terrain and an extension of the wildland fire that is not limited to the computational domain Ω under consideration, we assume horizontal isotropy. Even though an exact estimation of D is beyond the scope of the present study, a quantitative estimation of D is required to carry out sensitivity analysis related to turbulence and fire-spotting. D corresponds to the turbulent heat convection generated by the fire.

We shall adopt for such quantitave estimation the analytical representation whose derivation can be found in [48]. Thus, D will read

$$D \simeq 0.1\,\chi \left[\frac{\gamma\,\Delta T\,g\,h^3}{\nu\chi}\right]^{1/3} - \chi, \tag{9}$$

where χ is the air thermal diffusivity, γ is the thermal expansion coefficient, ΔT is the temperature difference in the convective cell, h is the dimension of the convective cell, g is the gravity constant and ν is the kinematic viscosity (see Table 2).

The selected parameterization for fire spotting as well is derived in [48]. Firebrand transport is characterized through the log-normal parameters μ and σ. μ describes firebrand lofting inside the convective column. The relative density and the atmospheric drag impact the buoyant forces acting on the firebrands; hence, it is appropriate to include these quantities in the definition of μ to describe the maximum allowable height for each firebrand for varying fireline intensity. The density ratio $\rho_\mathrm{a}/\rho_\mathrm{f}$ also limits the maximum allowable height for each firebrand. μ is thus defined as

$$\mu = H\left(\frac{3\,\rho_\mathrm{a}\,C_\mathrm{d}}{2\,\rho_\mathrm{f}^*}\right)^{1/2}, \tag{10}$$

where H [m] represents the plume height, which is related to the maximum loftable height H_p via the relation $H_p = \lambda\,H$, and where $\rho_f^* = \rho_f/\lambda^2$ [$\mathrm{kg\,m^{-3}}$] is the biomass fuel density that accounts for the correlation factor λ between smoke plume height and maximum allowable height for firebrands. We adopt the analytic formulation of H with respect to the fireline intensity I used in [80], i.e.

$$H = \alpha_H \, H_{\text{ABL}} + \beta_H \, \left(\frac{I}{dP_{\text{f0}}} \right)^{\gamma_H} \, \exp\left(\delta_H \, \frac{N_{\text{FT}}^2}{N_0^2} \right), \tag{11}$$

where α_H, β_H, γ_H and δ_H are empirical constant parameters, P_{f0} [W] is the reference fire power ($P_{\text{f0}} = 10^6 \, W$), H_{abl} [m] is the height of the atmospheric boundary layer (ABL), and the subscript FT stands for free troposphere.

The parameter σ characterizes the wind-aided transport of firebrands after they are ejected from the convective column. In a wind-driven regime of fire-spotting, the flight path of the firebrands is affected by their size, and firebrands beyond a critical size cannot be steered by the prevailing wind. This critical size is defined as the maximum liftable radius $r_{\text{max}} = ||\text{U}||^2/g$. It is interesting to note that the dimensionless ratio $||\text{U}||^2/(rg)$ (r is the brand radius) is also known as the Froude number: it quantifies the balance between inertial and gravitational forces applying on firebrands. So σ is computed as

$$\sigma = \frac{1}{2 z_{\text{p}}} \ln\left(\frac{\|\text{U}\|^2}{rg} \right). \tag{12}$$

Note that z_p corresponds to the pth percentile and can be estimated from the z-tables (`http://www.itl.nist.gov/div898/handbook/eda/section3/eda3671.htm`). We assume that the pth percentile represents the maximum landing distance for firebrands under different situations and no ignition is possible beyond this cut-off. The cut-off criteria is chosen empirically so that $z_p = 0.45$ as in [48], which corresponds to the 67th percentile point.

Statistical Description.. The following strategy is adopted to obtain a statistical description of these three parameters $\{D, \sigma, \mu\}$, which depend on a large set of subparameters.

The subparameters are perturbed around their nominal values found in the literature following uniform PDFs. To obtain a range of variation for D, we modify the parameters ΔT and h. As for parameters σ and μ, we modify the following parameters: α_H, β_H, γ_H , δ_H, H_{abl} in Eq. (11); ρ_{a}, ρ_{f} in Eq. (10); z_{p} and r in Eq. (12). For the parameters α_H, β_H, γ_H and δ_H, the extrema of the uniform PDF correspond to the highest and lowest values encountered in all the possible configurations described in [80]. ΔT varies in the range $[800; 923]$ K. For all other parameters, the extrema are defined such as adding a perturbation of 20 % to the values adopted in [48].

Once uniform PDFs are defined for each subparameter, we sample them through a Monte Carlo random sampling with sample size $N = 10,000$. Based on Eqs. (9)–(12), we thus obtain 10,000 realizations of the three parameters of interest $\{D, \sigma, \mu\}$. We can then analyze their empirical statistical distribution by fitting the resulting histograms with different types of PDF. Figure 1 presents the fits obtained when using a Beta-distribution for each sample. We adopt such distribution due to the requirement for positiveness, limitlessness, and compatibility with the available surrogates, in particular with the gPC given the Wiener–Askey scheme, see [81]. Table 5 presents the characteristics of each Beta-distribution and the associated range of variation for each parameter in $\boldsymbol{\theta} = (\mu, \sigma, D)^T$. We recall the analytic formulation for the Beta-distribution denoted by Beta, with a and b $(a, b > 0)$ the "shape parameters":

$$\mathrm{Beta}(\mathrm{x}; a, b) = \frac{\Gamma(a+b)\, \mathrm{x}^{a-1}\, (1-\mathrm{x})^{b-1}}{\Gamma(a)\Gamma(b)}, \tag{13}$$

for $\mathrm{x} \in (0,1)$, with $\Gamma(\mathrm{x})$ the Gamma function. To shift and/or scale the distribution, the "location" and "scale" parameters are introduced. More specifically, $\mathrm{Beta}(\mathrm{x}, a, b, \mathrm{location}, \mathrm{scale})$ is equivalent to $\mathrm{Beta}(\mathrm{y}, a, b)/\mathrm{scale}$ with $\mathrm{y} = (\mathrm{x} - \mathrm{location})/\mathrm{scale}$.

2.3. Simulated Quantities of Interest

We now define two scalar indices to represent the evolution of a fire over a time period $[0; T]$. We consider first the percentage of the computational domain Ω that is burnt at a given time t:

$$A_t = \frac{\int_\Omega \mathcal{I}_{\mathcal{B}(t)}(\mathrm{x}, t)\, d\mathrm{x}}{|\Omega|}, \tag{14}$$

where $|\Omega|$ [m^2] corresponds to the area of the computational domain and $\mathcal{I}_{\mathcal{B}(t)}$ is the indicator function of the burnt area, which returns 1 inside of the burnt area and 0 elsewhere. A_t corresponds to a normalized burnt area. However, this quantity does not give information on the topology of the fire, which can be complex in the case of fire-spotting. To overcome this limitation, we also consider an indicator S_t that describes the minimum spanning rectangle (MSR) of the burnt area over the area of the computational domain $|\Omega|$ at a given time:

$$S_t = \frac{|\mathrm{MSR}(t)|}{|\Omega|}. \tag{15}$$

The MSR is a geometrical quantity that corresponds to the smallest rectangle within which all burnt grid points lie at a given time t. So $|\mathrm{MSR}(t)|$ [m^2] measures the area of this rectangle. As an example, Fig. 2 presents an ensemble of 100 firelines at time 50 min, where each fireline corresponds to a different set of parameters D, μ and σ (i.e. a different realization of $\boldsymbol{\theta} = (\mu, \sigma, D)^T$)

Table 5: Range of variations and Beta-distribution for $\boldsymbol{\theta} = (\mu, \sigma, D)^T$. Note that the parameters of the Beta-distribution (Eq. 13) are given in the following order: shape parameters a and b, location and scale.

Parameter	Minimum/maximum values	Beta-distribution parameters
Log-normal parameter σ	5.49–12.69	1.37 1.99 5.94 4.93
Log-normal parameter μ	7.25–98.16	3.18 7.49 7.43 94.73
Turbulent diffusion coef. D $[\mathrm{m^2\,s^{-1}}]$	0.23–0.47	1.19 1.20 0.23 0.23

obtained by sampling the Beta-distributions given in Table 5. For each fireline, Fig. 2 shows the corresponding normalized MSR as defined in Eq. (15) at time 50 min. Low MSR values (rose colors) indicate simple topology of the fireline, while for high MSR values (yellow colors) the fireline presents more irregularities and a more complex propagation induced by turbulence and fire-spotting.

In this work, we analyze the time dependency of the quantities A_t and S_t by comparing them at two different times, $t_1 = 26$ min and $t_2 = 34$ min. The resulting scalar quantities (or "observables") are noted A_1, A_2, S_1 and S_2.

2.4. Numerical Implementation

The code `LSFire+` is developed in C and Fortran, where the turbulence and fire-spotting parametrization routines, labeled as `RandomFront 2.3b`, act as a post-processing routine at each time step in a level-set-method (LSM) code for the front propagation implemented through the library `LSMLIB` [83] and the ROS is computed by using the library `FireLib` [84]. The numerical library `LSMLIB` is written in Fortran2008/OpenMP. It advects the fireline through standard algorithms for the LSM, including also fast marching method algorithms. The aforementioned routines are freely available at the official git repository of BCAM, Bilbao, `https://gitlab.bcamath.org/atrucchia/randomfront-wrfsfire-lsfire`.

3. Surrogate Modeling

3.1. Principles and Notations

The objective of the present paper is to build surrogate models (or "response surfaces") that represent how the normalized burnt area A_t or the normalized MSR S_t (the generic scalar output is noted $\mathrm{y} \in \mathbb{R}$) changes with respect to a selection of the most relevant input parameters (the set of uncertain parameters is noted $\boldsymbol{\theta} \in \mathbb{R}^d$). The input stochastic space is defined either by $\boldsymbol{\theta} = (U, I, D)^T$ or $\boldsymbol{\theta} = (\mu, \sigma, D)^T$ (see Sec. 2.2); the size of the input stochastic space is $d = 3$.

The key idea of a surrogate is to replace the fire spread model $\mathrm{y} = \mathcal{M}(\boldsymbol{\theta})$ by a weighted finite sum of basis functions that can be generally expressed as

$$\widehat{\mathrm{y}}\,(\boldsymbol{\theta}) = \sum_{\boldsymbol{\alpha} \in \mathcal{A}} \gamma_{\boldsymbol{\alpha}}\, \Psi_{\boldsymbol{\alpha}}\,(\boldsymbol{\theta})\,, \tag{16}$$

where the coefficients $\gamma_{\boldsymbol{\alpha}}$ and the basis functions $\Psi_{\boldsymbol{\alpha}}$ are to be determined, $\mathcal{A}$ being the set of indices that defines the basis size. In practice, the coefficients and basis functions are calibrated by the training set (or "database") $\mathcal{D}_N$ that corresponds to a limited number N of forward model integrations (or "training set") such that

$$\mathcal{D}_N = (\Theta, \mathcal{Y}) = \left\{ \left(\boldsymbol{\theta}^{(k)}, \mathrm{y}^{(k)}\right)_{1 \leq k \leq N} \right\}, \tag{17}$$

where $\mathrm{y}^{(k)} = \mathcal{M}(\boldsymbol{\theta}^{(k)})$ corresponds to the integration of the forward model $\mathcal{M}$ (`LSFire+` in the present study) for the kth set of input parameters $\boldsymbol{\theta}^{(k)}$.

Two types of surrogate models are compared in the following: the gPC-expansion that retrieves the global forward model behavior on the one hand, the GP regression that is a local interpolator of the forward model behavior at the training points on the other hand. Different types of surrogate are tested to determine what is the best choice in the present application. For gPC-expansion, the user needs to determine the appropriate total polynomial order of the expansion as well as the appropriate type and number of basis functions $\Psi_{\boldsymbol{\alpha}}$. There are also different projection strategies to compute the coefficients $\gamma_{\boldsymbol{\alpha}}$. For GP regression, the user needs to choose the type of correlation structure and to estimate its associated hyperparameters.

3.2. Generalized Polynomial Chaos (gPC) Expansion

$\boldsymbol{\theta}$ is defined in the input physical space and its counterpart in the standard probabilistic space is noted $\boldsymbol{\zeta} = (\zeta_1, \cdots, \zeta_d)$, with ζ_i the random variable associated with the ith uncertain parameter θ_i in $\boldsymbol{\theta}$ characterized by its marginal PDF ρ_{θ_i}. $\boldsymbol{\theta}$ is thus rescaled in the standard probabilistic space to which the gPC framework applies.

3.2.1. Polynomial Basis

$\boldsymbol{\theta}$ is projected onto a stochastic space spanned by the orthonormal polynomial functions $\{\Psi_{\boldsymbol{\alpha}}(\boldsymbol{\zeta})\}_{\boldsymbol{\alpha}\in\mathcal{A}}$. The basis functions are orthonormal with respect to the joint PDF $\boldsymbol{\rho}_{\boldsymbol{\zeta}}(\boldsymbol{\zeta})$, i.e.

$$\langle \Psi_{\boldsymbol{\alpha}}(\boldsymbol{\zeta}), \Psi_{\boldsymbol{\beta}}(\boldsymbol{\zeta}) \rangle = \int_Z \Psi_{\boldsymbol{\alpha}}(\boldsymbol{\zeta})\, \Psi_{\boldsymbol{\beta}}(\boldsymbol{\zeta})\, \boldsymbol{\rho}_{\boldsymbol{\zeta}}\, d\boldsymbol{\zeta} = \delta_{\boldsymbol{\alpha}\boldsymbol{\beta}}, \tag{18}$$

with $\delta_{\boldsymbol{\alpha}\boldsymbol{\beta}}$ the Kronecker delta-function and $Z \subseteq \mathbb{R}^d$ the space in which $\boldsymbol{\zeta}$ evolves. In practice, the orthonormal basis is built using the tensor product of one-dimensional polynomial functions, $\Psi_{\boldsymbol{\alpha}} = \phi_{\alpha_1} \dots \phi_{\alpha_d}$ with ϕ_{α_i} the one-dimensional polynomial function. The choice for the basis functions depends on the probability measure of the random variables. According to Askey's scheme, the Jacobi polynomials form the optimal basis for random variables following Beta-distribution, and the Legendre polynomials are the counterpart for uniform distribution [81].

Assuming that the solution of the fire spread model is of finite variance, each quantity of interest y (see Sec. 2.3) can be considered as a random variable for which there exists a gPC expansion of the form

$$\widehat{\mathrm{y}}\,(\boldsymbol{\theta}) = \mathcal{M}_{\mathrm{pc}}(\boldsymbol{\theta}(\boldsymbol{\zeta})) = \sum_{\boldsymbol{\alpha}\in\mathcal{A}} \gamma_{\boldsymbol{\alpha}}\, \Psi_{\boldsymbol{\alpha}}\,(\boldsymbol{\zeta})\,. \tag{19}$$

$\Psi_{\boldsymbol{\alpha}}$ is the $\boldsymbol{\alpha}$th multivariate basis function chosen in adequacy with the PDF $\boldsymbol{\rho_\theta}$ associated with the parameters $\boldsymbol{\theta}$ (all random variables in $\boldsymbol{\theta}$ are assumed independent so that $\boldsymbol{\rho_\theta}$ is the product of the marginal PDFs $\{\rho_{\theta_i}\}_{i=1,\cdots,d}$). $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_d)$ is a multi-index in $\mathcal{A}$, which identifies the components of the multivariate polynomial $\Psi_{\boldsymbol{\alpha}}$.

Note that Eq. (19) represents how the normalized burnt area A_t or the normalized MSR S_t varies according to changes in the input vector $\boldsymbol{\theta}$. Once the PDF $\boldsymbol{\rho_\theta}$ is chosen, $\{\gamma_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathcal{A}}$ are the unknowns to determine to build the surrogate $\mathcal{M}_{\mathrm{pc}}$.

3.2.2. Truncation Strategy

For computational purposes, the sum in Eq. (19) is truncated to a finite number of terms r that is associated with the total polynomial order P of the gPC-expansion. There are several ways of choosing the number of terms r referred to as the "truncation strategy". Note that we will use the concept of "enumeration functions" in the following: a linear (or hyperbolic) enumeration function is a mapping $\mathfrak{J}$ from $\mathbb{N}$ to $\mathbb{N}^d$, which establishes a bijective mapping between a given integer i and a multi-index $\boldsymbol{\alpha}$.

Linear Truncation Strategy. The standard truncation strategy (referred to as "linear") consists in retaining in the gPC-expansion all polynomials involving the d random variables of total degree less or equal to P. Hence, $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_d) \in \{0, 1, \cdots, P\}^d$. The number of terms r is therefore constrained in this linear case by the number of random variables d and by the total polynomial order P so that

$$r_{\text{lin}} = \frac{(d+P)!}{d!\ P!}. \tag{20}$$

The set of selected multi-indices for the multi-variate polynomials $\mathcal{A}$ is defined as

$$\mathcal{A}_{\text{lin}} \equiv \mathcal{A}_{\text{lin}}(d, P) = \{\boldsymbol{\alpha} \in \mathbb{N}^d : |\boldsymbol{\alpha}| \leq P\} \subset \mathbb{N}^d, \tag{21}$$

where $|\boldsymbol{\alpha}| = ||\boldsymbol{\alpha}||_1 = \alpha_1 + \cdots + \alpha_d$ is the "total order" of the multi-index. In this case, we refer to the basis as the "full basis" for a given total polynomial order P.

Hyperbolic Truncation Strategy. As an alternative to the linear truncation strategy, the "hyperbolic" truncation strategy consists in eliminating a priori high-order interaction terms (i.e. polynomial terms involving more than one component of $\boldsymbol{\theta}$), see [70]. A more general way than Eq. (21) to define the number of terms r in the gPC expansion consists in introducing q-quasinorms:

$$\mathcal{A}_{\text{hyp}} \equiv \mathcal{A}_{\text{hyp}}(d, P, q) = \left\{ \boldsymbol{\alpha} \in \mathbb{N}^d : ||\boldsymbol{\alpha}||_q \leq P \right\}, \tag{22}$$

where the q-semi-norm is given by

$$||\boldsymbol{\alpha}||_q \equiv \left(\sum_{i=1}^{d} (\alpha_i)^q \right)^{1/q} . \tag{23}$$

The number of terms in the gPC-expansion is expressed by the cardinality of $\mathcal{A}$, which varies according to P and q for a fixed dimension d. The adoption of such semi-norms penalizes high-rank indices and high-order interactions. The lower the value of q, the higher the penalty in the determination of $\mathcal{A}$. When $q = 1$ we retrieve the linear truncation strategy and therefore a full basis of cardinality $\mathcal{A}_{\text{lin}}(d, P)$. In the following, we will study how the performance of the surrogate depends on the choice of the hyperbolic parameter $q \in [0, 1]$.

Sparse Truncation Strategies. There are alternatives to reduce the number of terms in the gPC-expansion. We will now schematically represent three of them, ordered by complexity: 1-"sequential strategy", 2- "cleaning strategy", 3- "least angle regression".

1- The sequential strategy [85] consists in constructing the gPC-expansion in an incremental way, starting from the first term Ψ_0 ($K_0 = \{0\}$) and adding one term at a time in the basis ($K_{i+1} = K_i \cup \{\Psi_{i+1}\}$). The terms that are sequentially added to the basis are ordered according to the adopted enumeration strategy (linear or hyperbolic). The response surface is therefore of increasing complexity, since the enumeration functions in both cases increase the polynomial complexity when increasing the index. In the present study, the construction process is stopped when a given accuracy is achieved, or when the number of terms in the gPC-expansion reaches the maximum size of the basis $r_{\max}$ specified by the user.

2- An alternative to the sequential strategy is the cleaning strategy [85], which builds a gPC-expansion containing at most $r_{\max}$ significant coefficients, i.e. at most $r_{\max}$ significant basis functions, starting from the full basis (still retaining the constraint of hyperbolic truncation if selected). The key idea of the cleaning strategy is to discard from the active basis the polynomials $\Psi_{\boldsymbol{\alpha}}$ that are associated with coefficients of low magnitude, i.e. satisfying

$$|\gamma_{\boldsymbol{\alpha}}| \leq \epsilon \cdot \max_{\boldsymbol{\alpha}' \in \mathcal{A}'} |\gamma_{\boldsymbol{\alpha}'}| \tag{24}$$

where ϵ is the significance factor set to 10^{-4}, and where $\mathcal{A}'$ represents the current active basis. This selection procedure means that the terms in the gPC-expansion are not ordered according to the degree of the polynomial functions but instead according to the magnitude of the coefficients.

3- In complement to the sequential and cleaning strategies, there is a more advanced approach called least-angle regression (LAR) to select the active polynomial terms. The key idea of the LAR approach is to select at each iteration a polynomial among the r terms of the full basis (or eventually the hyperbolic-truncated basis) based on the correlation of the polynomial term with the current residual. The selected term is added to the active set of polynomials. The coefficients of the active basis are computed so that every active polynomial is equicorrelated with the current residual until convergence is reached. Thus, LAR builds a collection of surrogates that are less and less sparse along the iterations. Iterations stop either when the full basis has been looked through or when the maximum size of the training set has been reached. When the iterations stopped, the polynomial coefficients are computed via the least-square algorithm presented below. More details can be found in [71, 70, 86].

3.2.3. Projection strategy

In this work, we focus on non-intrusive approaches based on ℓ_2-minimization methods to numerically compute the coefficients $\{\gamma_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathcal{A}}$ using the N snapshots from the training set $\mathcal{D}_N$.

Galerkin Pseudo-Spectral Projection. This Galerkin-type projection relies on the orthonormality property of the polynomial basis. Using this approach, the $\boldsymbol{\alpha}$th coefficient $\gamma_{\boldsymbol{\alpha}}$ is computed using the definition of the inner product that is numerically approximated using tensor-based Gauss quadrature (referred to as "quadrature" in the following) as follows

$$\gamma_{\boldsymbol{\alpha}} = \langle \mathrm{y}, \Psi_{\boldsymbol{\alpha}} \rangle \;\cong\; \sum_{k=1}^{N} \mathrm{y}^{(k)}\, \Psi_{\boldsymbol{\alpha}}(\boldsymbol{\zeta}^{(k)})\, w^{(k)}, \tag{25}$$

where $\mathrm{y}^{(k)} = \mathcal{M}(\boldsymbol{\theta}^{(k)})$ is the kth snapshot of the $\mathcal{D}_N$-database corresponding to the `LSfire+` simulation for the kth quadrature root $\boldsymbol{\theta}^{(k)}$ of $\Psi_{\boldsymbol{\alpha}}$, and where w^k is the weight associated with $\boldsymbol{\zeta}^{(k)}$ (corresponding to $\boldsymbol{\theta}^{(k)}$ in the standard probabilistic space). When considering a full basis, $(P+1)$ is the number of quadrature roots required in each uncertain direction to ensure an accurate calculation of the integral $\langle \mathrm{y}, \Psi_{\boldsymbol{\alpha}} \rangle$. Hence, in our problem, we have $N = (P+1)^3$ simulations in the training set to build the PC surrogates through

Galerkin pseudo-spectral projection.

Least-Square Minimization.. With this approach, the estimation of the coefficients $\{\gamma_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathcal{A}}$ is done by solving a least-square minimization problem, i.e. by minimizing the approximation error between the (exact) `LSfire+` model evaluations and the PC-surrogate estimations at the points of the training set $\mathcal{D}_N$. The least-square projection solves a minimization problem over the given basis as follows:

$$\widehat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}\in\mathbb{R}^r}{\operatorname{argmin}} \sum_{k=1}^{N} \left(\mathrm{y}^{(k)} - \sum_{\boldsymbol{\alpha}\in\mathcal{A}^P} \gamma_{\boldsymbol{\alpha}} \, \Psi_{\boldsymbol{\alpha}} \left(\mathbf{x}^{(k)}\right) \right)^2 \tag{26}$$

which is achieved through classical linear algebra algorithms. Note that the sample size N required by this strategy for the problem to be well posed is at least equal to $(r+1)$, where r is the number of gPC-coefficients (i.e. the cardinality of the set $\mathcal{A}$). Note also that least-square minimization is used here to compute the coefficients selected by the sparse truncation methods (sequential, cleaning or LAR). When using non-sparse truncation strategies, this projection method is referred to as the standard least-square (SLS) approach.

3.2.4. Workflow scheme for constructing the gPC-expansion

A complete algorithm relative to the implementation of the gPC-surrogate can be summarized as follows:

1. choose the polynomial basis $\{\Psi_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathcal{A}}$ according to the assumed marginal PDFs of the inputs $\boldsymbol{\theta} = (\|\mathrm{U}\|, I, D)^T$ or $\boldsymbol{\theta} = (\mu, \sigma, D)^T$;

2. choose the total polynomial degree P according to the complexity of the physical processes;
3. truncate the expansion to r_{lin} or r_{hyp} terms corresponding to the multi-index set $\mathcal{A}_{\text{lin}}$ or $\mathcal{A}_{\text{hyp}}$ using linear or hyperbolic truncation (r_{lin} depends on d, P; r_{hyp} depends on d, P and q with q the hyperbolic factor satisfying $0 < q \leq 1$);
4. in the case of a sparse strategy (sequential, cleaning or LAR), find a suitable set of multi-indices $\mathcal{A} \subset \mathcal{A}_{\text{lin,hyp}}$ with a cardinality $r \leq r_{\text{lin, hyp}}$, otherwise skip this step;
5. apply a projection strategy (quadrature or least-square) to compute the coefficients $\{\gamma_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha} \in \mathcal{A} \subset \mathbb{N}^d}$ using $N = (P+1)^d$ snapshots from the simulation database $\mathcal{D}_{N_{\text{ref}}}$;
6. formulate the surrogate model $\mathcal{M}_{\text{pc}}$, which can be evaluated for any new pair of parameters $\boldsymbol{\theta}^* = \left(\|\mathrm{U}\|^*, I^*, D^*\right)^T$ or $(\mu^*, \sigma^*, D^*)^T$.

3.3. Gaussian Process (GP) surrogate model

As stated by [67], a GP is a random process (here the observable from the fireline evolution y) indexed over a domain (here $\mathbb{R}^d$), for which any finite collection of process values (here $\left\{\mathrm{y}(\boldsymbol{\theta}^{(k)})\right\}_{1\leq k\leq N}, \boldsymbol{\theta}^{(k)} \in \Theta$) has a joint Gaussian distribution. Concretely, let $\widetilde{\mathrm{y}}$ be a Gaussian random process fully described by its zero mean and its correlation π:

$$\widetilde{\mathrm{y}}(\boldsymbol{\theta}) \sim \mathrm{GP}\left(0, \sigma_{\text{gp}}^2\, \pi(\boldsymbol{\theta}, \boldsymbol{\theta}')\right), \tag{27}$$

with $\pi(\boldsymbol{\theta}, \boldsymbol{\theta}') = \mathbb{E}\left[\widetilde{\mathrm{y}}(\boldsymbol{\theta})\widetilde{\mathrm{y}}(\boldsymbol{\theta}')\right]$. In the present case, the correlation function π (or kernel) is chosen as a squared exponential (also known as “RBF kernel”, RBF standing for radial basis function):

$$\pi(\boldsymbol{\theta}, \boldsymbol{\theta}') = \exp\left(-\frac{\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2}{2\,\ell_{\text{gp}}^2}\right), \tag{28}$$

where ℓ_{gp} is a length-scale representing the model output dependency between two inputs $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, and where σ_{gp}^2 is the variance of the observable. The surrogate model is thus the mean of the GP, resulting of conditioning $\widetilde{\text{y}}$ on the training set $\mathcal{Y} = \left\{\text{y}\left(\boldsymbol{\theta}^{(k)}\right)\right\}_{1\leq k\leq N}$. The quantity of interest provided by the GP-surrogate for any given $\boldsymbol{\theta}^* \in \mathbb{R}^d$ satisfies

$$\text{y}_{\text{gp}}(\boldsymbol{\theta}^*) = \sum_{k=1}^{N} \beta_k\, \pi\left(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(k)}\right), \tag{29}$$

where

$$\beta_k = \left(\boldsymbol{\Pi} + \tau_{\text{gp}}^2\, \mathbf{I}_N\right)^{-1} \left(\text{y}(\boldsymbol{\theta}^{(1)}) \ldots \text{y}(\boldsymbol{\theta}^{(N)})\right)^T, \tag{30}$$

$$\boldsymbol{\Pi} = \left(\pi(\boldsymbol{\theta}^{(j)}, \boldsymbol{\theta}^{(k)})\right)_{1\leq j,k\leq N}, \tag{31}$$

and where τ_{gp} (referred to as the "nugget effect") is used to avoid ill-conditioning issues for the matrix $\boldsymbol{\Pi}$. The hyperparameters $\{\ell_{\text{gp}}, \sigma_{\text{gp}}, \tau_{\text{gp}}\}$ are optimized through maximum likelihood applied to the dataset $\mathcal{D}_N$ using a basin hopping technique [87].

3.4. Design of Experiments

We build several datasets to analyze the performance of the gPC- and GP-surrogates in an extensive way in Section 4; these datasets are summarized in Table 6. Note that estimating the generalization error of the surrogate model requires the use of an independent dataset, that is why we use a Monte Carlo random sampling including $N = 216$ members for validation. Note also that the Halton's low-discrepancy sequence is involved in this work in order to explore the hypercube defined by the distribution of the uncertain parameters. This design of experiment will be compared to a tensor-based Gauss quadrature in terms of performance of the surrogate model. The reader shall refer to Section 2.2 for more details on the range of variation and the marginal PDFs of each uncertain parameter.

Table 6: Datasets $\mathcal{D}_N$ of `LSfire+` simulations used in this work for building surrogates ("training") or for validating them ("validation").

Sampling Strategy	Purpose	Sample size
	$\boldsymbol{\theta} = (\lVert \mathrm{U} \rVert, I, D)^T$	
Halton's sequence	Training	216
Monte Carlo random sampling	Validation	216
	$\boldsymbol{\theta} = (\mu, \sigma, D)^T$	
Halton's sequence	Training	216
Quadrature rule	Training	216
Monte Carlo random sampling	Validation	216

3.5. Error Metrics

In the present study, two error metrics are used to assess the quality of the surrogate predictions: the empirical error between the surrogate prediction and the `LSfire+` model prediction (also known as "training error") on the one hand, and the Q_2 predictive coefficient [55] on the other hand.

3.5.1. Empirical Error ϵ_{emp}

The truncation of the gPC-expansion can eventually introduce an approximation error at the training points, which can be computed posterior to the surrogate construction. This empirical error denoted by ϵ_{emp} reads

$$\epsilon_{\text{emp}} = \frac{1}{N} \sum_{k=1}^{N} \left(\mathrm{y}^{(k)} - \widehat{\mathrm{y}}^{(k)} \right), \tag{32}$$

with $\mathrm{y}^{(k)}$ the kth element of the training set $\mathcal{D}_N$ (either the Halton's low discrepancy sequence or the quadrature database, see Table 6) and $\widehat{\mathrm{y}}^{(k)}$ the corresponding value predicted by the surrogate for the same element of the training set.

However, this error estimator has several drawbacks. First, the GP-model (built without noise in the kernel) is an interpolator so that the approximation error is expected to be $\epsilon_{\text{emp}} = 0$. Second, this estimator may severely underestimate the magnitude of the mean square error. When the size of the training set N comes closer to the cardinality of the gPC-expansion $\mathcal{A}$, ϵ_{emp} may tend to zero, while the actual mean square error does not; this issue is known as "overfitting".

3.5.2. Predictive coefficient Q_2

We require a more robust error estimator suitable for both gPC-expansion and GP-model. In this work, we use the Q_2 predictive coefficient based on cross-validation. The computation of Q_2 relies on two distinct datasets: the current training set $\mathcal{D}_N$ (either the Halton's sequence or the quadrature database) and a Monte Carlo sample $\mathcal{D}_{N_{\text{ref}}}$ that is independent of the surrogate construction and that is therefore referred to as the "validation dataset".

Q_2 is computed as

$$Q_2 = 1 - \frac{\sum_{k=1}^{N_{\text{ref}}} \left(\text{y}^{(k)} - \widehat{\text{y}}^{(k)}\right)^2}{\sum_{k=1}^{N_{\text{ref}}} \left(\text{y}^{(k)} - \overline{\text{y}}_{\text{ref}}\right)^2}, \tag{33}$$

with $\text{y}^{(k)}$ the kth element of the Monte Carlo sample $\mathcal{D}_{N_{\text{ref}}}$, $\widehat{\text{y}}^{(k)}$ the surrogate prediction for the same element of $\mathcal{D}_{N_{\text{ref}}}$ and $\overline{\text{y}}_{\text{ref}}$ the empirical mean over the Monte Carlo sample $\mathcal{D}_{N_{\text{ref}}}$. Note that computing Q_2, the training set $\mathcal{D}_N$ is only used to construct the surrogate model and to obtain the estimation $\widehat{\text{y}}$ of the quantity of interest y. The target value for Q_2 is 1.

3.6. Statistical Analysis

Once the surrogates are available for the different observables (A_1, A_2, S_1, S_2 – see Section 2.3), the statistics of the quantities of interest can be obtained. For the gPC-expansion, they can be derived analytically from the coefficients. For the GP-surrogate, we evaluate the surrogate predictions over a new dataset $\mathcal{D}_{N_{\text{sample}}}$ of size $N_{\text{sample}} = 10,000$ that is a subset of $\mathbb{R}^3$ and that is obtained using a standard Monte Carlo random sampling; this dataset is only used as input to the surrogate model and not to `LSfire+`.

3.6.1. Estimation of Statistical Moments

The mean value and STD of the observable y can be estimated as

$$\mu_{\widehat{\mathrm{y}}} = \frac{1}{N_{\text{sample}}} \sum_{k=1}^{N_{\text{sample}}} \widehat{\mathrm{y}}^{(k)}, \tag{34}$$

$$\sigma_{\widehat{\mathrm{y}}} = \sqrt{\frac{1}{N_{\text{sample}} - 1} \sum_{k=1}^{N_{\text{sample}}} \left(\widehat{\mathrm{y}}^{(k)} - \mu_{\widehat{\mathrm{y}}}\right)^2}, \tag{35}$$

with $\widehat{\mathrm{y}}^{(k)}$ the kth element of the dataset $\mathcal{D}_{N_{\text{sample}}}$ containing the surrogate evaluations over the aforementioned Monte Carlo sampled points.

Using the gPC-surrogate, the statistical moments can be derived analytically from the coefficients $\{\gamma_{\boldsymbol{\alpha}}\}_{\boldsymbol{\alpha}\in\mathcal{A}\subset\mathbb{N}^d}$ such that the mean and the STD read:

$$\mu_{\widehat{\mathrm{y}}_{\text{pc}}} = \gamma_0, \tag{36}$$

$$\sigma_{\widehat{\mathrm{y}}_{\text{pc}}} = \sqrt{\sum_{\substack{\boldsymbol{\alpha}\in\mathcal{A}\subset\mathbb{N}^d \\ \boldsymbol{\alpha}\neq 0}} \gamma_{\boldsymbol{\alpha}}^2}. \tag{37}$$

3.6.2. Sensitivity Analysis Diagnostics

Sobol' indices [53, 49] are commonly used for sensitivity analysis based on variance analysis. They provide the quantification of how much of the variance in the quantity of interest is due to the variance in the input parameters assuming (1) these input random variables are independent and (2) the random output is squared integrable.

The variance of an output random variable y denoted by $\mathbb{V}[y]$ can be decomposed as

$$\mathbb{V}[y] = \sum_{i=1}^{d} \mathbb{V}_i(y) + \sum_{j=i+1}^{d} \mathbb{V}_{ij}(y) + \cdots + \mathbb{V}_{1,2,\ldots,d}(y), \tag{38}$$

where $\mathbb{V}_i(y) = \mathbb{V}\left[\mathbb{E}(y|\theta_i)\right]$, $\mathbb{V}_{ij}(y) = \mathbb{V}\left[\mathbb{E}(y|\theta_i, \theta_j)\right] - \mathbb{V}_i(y) - \mathbb{V}_j(y)$ and more generally,

$$\mathbb{V}_I(y) = \mathbb{V}\left[\mathbb{E}(y|\theta_I)\right] - \sum_{J \subset I \text{ s.t. } J \neq I} \mathbb{V}_J(y),\ \forall I \subset \{1, \ldots, d\} \tag{39}$$

Based on this variance decomposition, the first-order Sobol' index S_i associated with the ith parameter of $\boldsymbol{\theta}$ is given by

$$S_i = \frac{\mathbb{V}_i(y)}{\mathbb{V}(y)}, \tag{40}$$

and corresponds to the ratio of the output variance $\mathbb{V}(y)$ that is uniquely related to the ith input parameter; S_i ranges between 0 and 1. The corresponding total Sobol' index S_{T_i} measures the whole contribution of the ith input parameter (including *interactions with other parameters* of $\boldsymbol{\theta}$) on the output variance. Its definition reads

$$S_{T_i} = \sum_{\substack{I \subset \{1,\ldots,d\} \\ I \ni i}} S_I. \tag{41}$$

By definition, $S_{T_i} \geq S_i$. If both first-order and total indices are not equal, this means that the input parameter θ_i share some interactions with other parameters of $\boldsymbol{\theta}$.

For the GP-surrogate approach, Sobol' indices are stochastically estimated using Martinez' formulation since this estimator is stable and provides asymptotic confidence intervals for first-order and total-order indices [88].

For the gPC-expansion approach, Sobol' indices can be directly derived from the gPC-coefficients. For the ith component of the input random variable $\boldsymbol{\theta}$, the Sobol' index $\mathbb{S}_{\mathrm{pc},i}$ reads:

$$\mathbb{S}_{\mathrm{pc},i} = \frac{1}{(\sigma_{\widehat{\mathrm{y}}_{\mathrm{pc}}})^2} \sum_{\substack{\boldsymbol{\alpha} \in \mathcal{A}_i \subset \mathbb{N}^d \\ \boldsymbol{\alpha} \neq 0}} (\gamma_{\boldsymbol{\alpha}})^2 \,, \tag{42}$$

where $\sigma_{\widehat{\mathrm{y}}_{\mathrm{pc}}}$ is the STD computed in Eq. (37), and where $\mathcal{A}_i$ is the set of multi-indices selected in $\mathcal{A}$ such that the computation of $\mathbb{S}_{\mathrm{pc},i}$ only includes terms that depend on the input variable θ_i, namely

$$\mathcal{A}_i = \{\boldsymbol{\alpha} \in \mathbb{N}^d, |\boldsymbol{\alpha}| \leq P \mid \alpha_i > 0, \alpha_{k \neq i} = 0\}. \tag{43}$$

3.7. Numerical Implementation

The GP implementation relies on the Python package *scikit-learn* [89] (see http://scikit-learn.org/). The gPC-implementation relies on the Python package *OpenTURNS* [85] (see www.openturns.org). The *batman* [90] Python package is used to build datasets and perform statistical analysis.

4. Results

The objective of this study is two-fold. First, we provide an extensive comparison of the performance of different surrogate strategies (see Table 7) for a given training set $\mathcal{D}_N$; We evaluate their impact on the predicted quantities of interest A_t and S_t in terms of mean value and STD, but also their impact on the predicted Sobol' sensitivity indices. This extensive analysis is carried out for the case $\boldsymbol{\theta} = (\mu, \sigma, D)^T$, related to the fluctuating part of the model. Second, we use this framework to rank the uncertain parameters, either $\boldsymbol{\theta} = (\|\mathrm{U}\|, I, \tau)^T$ or $\boldsymbol{\theta} = (\mu, \sigma, D)^T$, by order of importance and identify the most influential input parameters.

4.1. Comparison of surrogate performance

4.1.1. Error assessment

Table 8 presents the error metrics (i.e. the ϵ_{emp} empirical error and the Q_2 predictive coefficient) obtained for different types of surrogate (gPC on the one hand, and GP on the other hand) with respect to $\boldsymbol{\theta} = (\mu, \sigma, D)^T$ but for a given size of the training set $N = 216$. The performance of the gPC-surrogate is analyzed in details for varying truncation and projection schemes summarized in Table 7; the GP-surrogate is obtained using a standard RBF kernel and is considered here as a basis for comparison in order to evaluate the quality of the gPC-surrogates. For each approach, one surrogate model is built for each of the four observables $\{A_1, A_2, S_1, S_2\}$ corresponding to the two quantities of interest A_t and S_t at times $t_1 = 26$ min and $t_2 = 34$ min.

In Table 8 we first focus on the results obtained with linear truncation ($q = 1$), meaning that the basis of polynomial functions is full for a given total polynomial order P. Figure 3 (right figure of each pair) presents corresponding scatter plots (referred to as "adequacy plots") of the surrogate predictions with respect to the physical model predictions. These plots quantify the adequacy of the surrogate to the physical model at the training points in terms of predicted burnt area ratio A_2. It is found that the Q_2 predictive coefficient is over 0.9 only for the LAR and cleaning sparse methods for all observables. The empirical error is of the same order of magnitude, varying between 10^{-3} for the MSR ratio S_t and 10^{-4} for the burnt area ratio A_t. Note that for a given observable at a given time, there is no significant difference among the surrogate strategies in terms of empirical error. We therefore focus the following analysis on the standalone Q_2 predictive coefficient. Note also that the performance of each surrogate is time independent since for a given observable, the Q_2 predictive coefficient is similar at times t_1 and t_2. We therefore focus on results at time t_2 in the following.

When moving to hyperbolic truncation schemes ($q = 0.75$ or $q = 0.5$), we reduce a priori the number of coefficients to compute in the gPC-expansion, while the size of the training set remains the same ($N = 216$). The lower the value of q, the smaller the number of gPC-coefficients r. Figure 4 (right plot of each pair) presents adequacy plots for hyperbolic truncation with $q = 0.5$; this is to compare to the adequacy plots obtained for linear truncation in Figure 3 (right plot of each pair). Results show that the performance of the quadrature approach does not improve when q decreases. In the opposite, the performance of the SLS approach improves and features a Q_2 predictive coefficient over 0.9 for A_2 and over 0.8 for S_2 when using hyperbolic truncation. This improvement is also noticeable in Figure 4 (right plot of each pair), where hyperbolic truncation allows to better represent the model response for low values of the burnt area ratio ($A_2 < 0.03$). The sequential sparse method also provides better results for a hyperbolic coefficient $q = 0.5$. The performance of LAR and cleaning sparse methods remains similar as in the linear case $q = 1$.

LAR appears as the most accurate gPC strategy and has a Q_2 predictive coefficient that is similar to that obtained with the GP-model based on RBF kernel. Hyperbolic truncation does not add much value to the results compared to linear truncation, except for the SLS strategy. This may be explained by the fact that the terms that are important to retain in the gPC-expansion are not located in an isotropic way in the three dimensions ($d = 3$). It is therefore of interest to identify which polynomial terms are important to keep in the basis in order to obtain a good performance of the surrogate in each of the three dimensions.

4.1.2. Sensitivity of gPC-surrogates to total polynomial order P

In Table 8, the results for SLS and LAR methods are obtained by choosing the optimal value of the total polynomial order P in the sense that the surrogate was obtained by finding the value of P that maximizes the Q_2 predictive coefficient; P varying between 1 and 14. Recall that the total polynomial order P determines the size of the full basis used to construct the surrogate when using linear truncation. The SLS method considers the full basis, while the LAR method selects the most influential terms among the full basis. Since the size of the training set is fixed to $N = 216$ and since $(P + 1)^3 = 216$ for $P = 5$, we know that the problem becomes ill-posed for a full basis when the total polynomial order is over 5. This is not an issue for LAR since it selects inline the influential coefficients in the basis. It is therefore of interest to investigate if the LAR method features an improved performance when $P > 5$.

Figure 5 presents the Q_2 predictive coefficient for P varying between 1 and 14 for SLS and LAR surrogates obtained for the burnt area ratio A_2. As expected, Fig. 5a shows that the best performance of the SLS method with linear truncation is obtained for $P = 5$ and that it degrades very fast when increasing P (the Q_2 predictive coefficient is below 0.4 for $P > 6$). When moving to hyperbolic truncation with $q = 0.5$, Fig. 5c shows that the Q_2 predictive coefficient remains over 0.4 for $P > 5$. The resulting surrogate is therefore improved in this configuration as already pointed out in Table 8. Hyperbolic truncation allows the SLS approach to include high-order polynomials in the basis without generating an ill-posed problem (i.e. without having more coefficients to compute than the size N of the training set). Still, results show that the Q_2 predictive coefficient does not follow a monotonically increasing function toward the target value 1 in this hyperbolic configuration; this configuration is therefore not robust. In the opposite, the LAR method shows a monotonic convergence towards the target value 1 when increasing P in Figs. 5b–d. A good performance of LAR is obtained for $P = 10$ for both linear and hyperbolic truncation schemes.

This sensitivity study shows that a total polynomial order P higher than 5 is required to build the response surface of the burnt area ratio. Similar results are obtained for the MSR ratio (not shown here). This demonstrates the benefits from sparse schemes when having a fixed and limited training set $\mathcal{D}_N$. Improving the performance of the SLS approach using linear truncation would require a higher total polynomial order P and therefore a larger training set.

4.1.3. Identification of the influential gPC-coefficients

Figure 3 (left figure of each pair) presents a three-dimensional schematic (referred to as "sparsity plot") of the coefficients retained in the gPC-expansion using linear truncation, each dimension corresponding to one stochastic/uncertain dimension. The three dimensions are here the turbulent diffusion coefficient D and the lognormal parameters μ and σ.

Quadrature and SLS methods have the same full basis for a given polynomial order P (here $P = 5$ since the size of the training set is $N = 216$); they are associated with a typical "pyramidal" sparsity plot, where the first coefficient corresponding to the mean estimate of the burnt area ratio A_2 has the highest magnitude (approximately equal to 0.04). For sparse methods (LAR, cleaning, sequential), the number of coefficients is significantly reduced since the terms with the least impact are automatically filtered out of the sparse basis. The sparsity plot has no longer a "pyramidal" shape. LAR and sequential strategies feature instead a two-dimensional structure (along the vertical plane) indicating that the burnt area ratio A_2 is not sensitive to the third dimension, here the lognormal parameter μ, but only to the lognormal parameter σ and to the turbulent diffusion coefficient D. Only the cleaning strategy retains a three-dimensional structure by accounting for interaction terms involving the lognormal parameter μ. This highlights the presence of influential interaction terms involving several parameters. However, all sparse strategies indicate that one direction is dominant since the number of coefficients in this direction is high and the basis terms can go up to a total polynomial order $P = 12$ in the case of cleaning and $P = 8$ in the case of LAR (instead of the constrained $P = 5$ for quadrature and SLS). This dominant direction corresponds to the lognormal parameter σ.

Note that Figure 4 (left figure of each pair) presents similar plots as Figure 3 (left figure of each pair) but for hyperbolic truncation with $q = 0.5$. The magnitude of the coefficients does not change for quadrature, explaining why hyperbolicity does not improve the performance of the surrogate based on quadrature. This is not the case of SLS, which now features high magnitude for the coefficients along the direction D for polynomial terms having a degree between 4 and 8. This highlights the need to have polynomials of higher degree to capture underlying physical processes. Still, SLS with hyperbolicity is not sufficient to capture the same structure as sparse methods. Note that sparse methods converge to the same structure using linear or hyperbolic truncation schemes, indicating the robustness of these methods.

The influence of the three parameters on the behavior of the burnt area ratio A_2 can be quantified using Sobol' sensitivity indices. Table 9 presents the Sobol' indices using sparse methods and linear truncation for the burnt area ratio A_2 (same results are obtained using hyperbolic truncation with $q = 0.5$ – not shown here). Table 10 presents similar quantities for the MSR ratio S_2. Results confirm that the lognormal parameter σ is the most influential one for both quantities of interest A_2 and S_2 with a first-order sensitivity index above 0.98 for A_2 and above 0.92 for S_2. This means that more than 90 % of the variance in A_2 and S_2 is explained by uncertainties in the lognormal parameter σ. Results also show interaction effects are limited but still present between the lognormal parameter σ and the turbulent diffusion parameter D as foreseen in sparsity plots. Note that all sparse gPC-surrogates as well as the GP-model exhibit the same global trend. The main differences lie in the relevance of the lognormal parameter μ. LAR and sequential strategies cut out any contribution of μ in the variability of the predicted quantities of interest. This is not the case of the cleaning strategy that has a non-zero total Sobol' index for μ as the GP-model.

We can evaluate the impact of the choice in the surrogate strategy on the predicted mean and STD estimates of the quantities of interest. Table 11 presents the mean and STD estimate of the burnt area ratio A_2 and of the MSR ratio S_2 obtained for different gPC- and GP-surrogates. Results show the consistency of the statistical moments obtained using sparse gPC-expansions and GP-model for both A_2 and S_2. The SLS approach using linear truncation is able to retrieve accurate mean and STD estimates (about 1 % deviation with respect to GP-model predictions). In the opposite, the quadrature approach provides mean and STD estimates with more than 10 % deviation with respect to GP-model predictions.

This highlights the importance of having high-order polynomial terms in some uncertain directions to build an accurate gPC-expansion and have accurate estimate of the statistical moments in the present study. These directions can be identified using Sobol' sensitivity indices. Sparse gPC-strategies are relevant to address such issues due to the flexibility of selecting the most influential polynomial terms during the construction of the surrogate (linear and hyperbolic schemes are defined a priori).

4.1.4. Sensitivity to the size of the training set

So far the analysis was obtained for a fixed training set of size $N = 216$ (generated using Halton's low discrepancy sequence or tensor-based Gauss quadrature in the case of quadrature). It is of interest to study if the same level of accuracy could be obtained for sparse gPC-surrogates built with a reduced training set ($N < 216$). To answer this question, we provide a convergence test for a training size N varying between 10 and 216 with respect to the observable S_2. For each size of the training set, a LAR gPC-surrogate is built and cross-validated using the available Monte Carlo database (Table 3.4) through the computation of the Q_2 predictive coefficient. We carry out this convergence test for different truncation strategies, i.e. for different levels of hyperbolicity $q \in \{1, 0.75, 0.5\}$. Figure 6 presents the evolution of Q_2 with respect to the size of the training set N. Results show the convergence of Q_2 to a constant value for $N > 100$. Linear truncation and hyperbolic truncation ($q = 0.5$) provide similar performance for $N > 100$. As before, we note that the hyperbolic solution obtained using $q = 0.75$ is not the best option.

4.2. Analysis of the physical model predictions

Results show that the LAR gPC-strategy features a good performance. In the following, we will use this strategy to further analyze the fire-spotting and turbulence submodel included in `LSFire+`. We summarize in Table 13 and Table 15 the error metrics as well as the mean and STD estimate of the burnt area ratio A_2 and of the MSR ratio S_2 at time t_2 for the two sets of uncertain parameters $\boldsymbol{\theta} = (\|\mathrm{U}\|, I, \tau)^T$ and $\boldsymbol{\theta} = (\mu, \sigma, D)^T$, respectively. Table 12 and Table 14 present the corresponding Sobol' Indices. Note that the following analysis holds for any time t since we show that results can be considered as time-independent. Note also that the empirical error ϵ_{emp} and the Q_2 predictive coefficient are in acceptable range for all tested configurations; we focus here on the physics of the problem.

Sobol' sensitivity indices order by relevance each parameter. In the case $\boldsymbol{\theta} = (\|\mathrm{U}\|, I, \tau)^T$, a clear dominance of the wind speed $\|\mathrm{U}\|$ is observed for the considered range of the fireline intensity I. This is a rather interesting result, since the normalization performed on the ROS model (i.e. parameter

α_{w} in Eq. 7) makes the propagation of the deterministic fireline depending solely on the orientation of the wind vector and not on its magnitude. This means that the wind has a more general and fundamental role as reflected also in the enhancement of fire-spotting and secondary fire generation.

The ballistic term σ in Eq. (5) strongly depends on the value of $\|\mathrm{U}\|$. This is in line with the results of the second set of input parameters. In the case $\boldsymbol{\theta} = (\mu, \sigma, D)^T$, σ is the most influential parameter when considering Sobol' indices, far above D and μ (in order of relevance). The trend for the observables A_t and S_t is comparable, still S_t gives slightly more relevance to μ and D inputs than A_t. As expected, for both parameter sets, the mean of the S_2-observable is larger than that of A_2. Its STD is also larger. Uncertainties in $\{\|\mathrm{U}\|, I, \tau\}$ induce a more significant spread of the fireline position and shape compared to uncertainties in $\{\mu, \sigma, D\}$. This is due to the fact that in the first case we also vary the ember ignition time scale.

In summary, these results highlight the importance of the mean wind factor, on the main fire propagation but also on the generation of secondary fires. This is consistent with the phenomenology of wildland fires and with the process of fire-spotting. In particular, fire-spotting refers to independent ignitions located far away from the main fireline. This process is accounted in the model via the lognormal parameter σ. The importance of σ is a proper mathematical feature of the adopted lognormal PDF for firebrand landing distance, since it controls the tail of the density function, the kurtosis of the lognormal density being equal to $\mathrm{e}^{4\sigma^2} + 2\mathrm{e}^{3\sigma^2} + 3\mathrm{e}^{2\sigma^2} - 3$. Hence this study shows that the new submodel correctly includes the double role of the mean wind, enhancing the propagation of the main fireline on the one hand, and carrying away firebrands for secondary ignitions on the other hand.

5. Discussion and Conclusions

This study presents an extensive comparative study of surrogate approaches to the nonlinear and multi-scale problem of turbulence and fire-spotting in wildland fire modeling, fire-spotting being a random process in which firebrand generation, emission and landing distance are intrinsically governed by the fire strength. A surrogate modeling approach is useful to analyze in a cost-effective way, how the fireline position and topology change according to variations in the input parameters for the new physical submodel introduced by Pagnini et al [45, 46, 47, 48] based on a randomized representation of the fireline. Results are presented from both algorithmic and physical perspectives. From an algorithmic viewpoint, it is of interest to compare several approaches to carry out global sensitivity analysis and to select which ones are accurate and computationally efficient. From a wildland fire perspective, uncertainty quantification and sensitivity analysis is a good practice to analyze any new submodel, spot unimportant parameters and identify which parameters are dominant for obtaining a good representation of turbulence and fire-spotting.

In this work, fast surrogate models based on generalized Polynomial Chaos (gPC) and Gaussian Process (GP) were used to limit the required number of physical model evaluations to at least 100. We analyzed the performance of different formulations of the gPC-surrogate in terms of design of experiments (how to choose the training points? how many training points are required to achieve a certain accuracy?), polynomial basis structures (how to select the influential terms of the polynomial basis?) and projection schemes (how to compute the coefficients of the gPC-expansion?). The generalization error of these surrogates was classically estimated using the Q_2 predictive coefficient. Sparse gPC-methods have shown their accuracy in line with the GP model based on RBF kernel, but with a less cumbersome representation for Sobol' indices and statistical moments. Sparse methods provide more flexibility to select high-order polynomial terms in a given direction of the uncertain space, without requiring more physical model evaluations and therefore without increasing the computational cost of sensitivity analysis. The best performance for the gPC-surrogate was obtained using a sparse least-angle regression (LAR) with a training set built using a Halton's low discrepancy sequence. Using this approach, the new parametrization `RandomFront 2.3b` for turbulence and fire-spotting was found to be a non-linear model with a remarkable range of variations in the size and topology of the fire due to uncertainties in its input parameters. There is a clear dominance of the lognormal parameter σ characterizing firebrand downwind transport and of the wind magnitude $\|\mathrm{U}\|$, which confirms that fire-spotting is a wind-driven, ballistic phenomenon.

Several issues can be met when building a robust surrogate model. First, when the problem is multi-scale, i.e. when uncertain parameters have correlation length-scales differing by several order of magnitudes. Sparse methods may filter out the less influential parameters. The LAR-based gPC surrogate was found to filter out the information coming from parameters with large length-scale. The cleaning-based surrogate proved to preserve these information, which may be important in a multi-scale problem such as fire-spotting.

Second, when choosing how to sample the stochastic space and construct the training set. Standard projection schemes such as tensor-grid Gauss quadrature and standard least-square methods have shown their limitations: a large part of the training set was wasted in regions of the parameter space far from the nonlinear processes to be explored. In the opposite, sparse methods based on least-square projection were found to identify in which stochastic direction the physical processes are more complex and require higher order polynomials or high-order interaction terms. Using hyperbolic truncation was not flexible enough for this purpose.

The increasing strength and occurrence of megafires due to climate change calls for the development of new tools for the prediction of fire occurrence, growth and frequency at regional scales. Reliable wildland fire spread models are a promising approach to provide short-term variability of fire danger. Statistical methods such as uncertainty quantification and sensitivity analysis also have an important role to play [91, 92, 93]. Present work pushes toward the integration of fire-spotting into regional-scale operational wildland fire spread simulators. This is the main direction of the future developments of this research. Future work will also include the extension of the surrogate approaches to vectorial inputs and outputs, in order to analyze the sensitivity of the fire behavior to a wind field and to describe the fire situation as a map instead of a scalar variable.

Acknowledgements. This research is supported by the Basque Government through the BERC 2014–2017 and BERC 2018–2021 programs, by the Spanish Ministry of Economy and Competitiveness MINECO through BCAM Severo Ochoa accreditations SEV-2013-0323 and SEV-2017-0718 and through project MTM2016-76016-R ”MIP”, and by the PhD grant ”La Caixa 2014”. The authors acknowledge EDF R&D for their support on the OpenTURNS library. They also acknowledge Pamphile Roy and Matthias De Lozzo at CERFACS for helpful discussions on *batman* and *scikit-learn* tools.

References

[1] D. Viegas, Forest Fire Propagation, Philosophical Transactions: Mathematical, Physical and Engineering Sciences 356 (1998) 2907–2928.

[2] R. Linn, J. Reisner, J. Colman, J. Winterkamp, Studying wildfire behavior using FIRETEC, Int. J. Wildland Fire 11 (2002) 233–246.

[3] W. Mell, M. Jenkins, J. Gould, P. Cheney, A physics-based approach to modelling grassland fires, Int. J. Wildland Fire 16 (2007) 1–22.

[4] S. Strada, C. Mari, J. B. Filippi, F. Bosseur, Wildfire and the atmosphere: Modelling the chemical and dynamic interactions at the regional scale, Atmos. Environ. 51 (2012) 234–249.

[5] M. Finney, J. Cohen, S. McAllister, W. Matt Jolly, On the need for a theory of wildland fire spread, Int. J. Wildland Fire 22 (1) (2013) 25–36.

[6] S. McAllister, M. Finney, Convection ignition of live forest fuels, Fire Safety Science 11 (2014) 1312 – 1325. `doi:10.3801/IAFSS.FSS.11-1312`.

[7] T. L. Clark, M. A. Jenkins, J. Coen, D. Packham, A coupled atmospheric-fire model: convective feedback on fire-line dynamics, J. Appl. Meteor. 35 (1996) 875–901.

[8] B. E. Potter, A dynamics based view of atmosphere-fire interactions, Int. J. Wildland Fire 11 (2002) 247–255.

[9] B. E. Potter, Atmospheric interactions with wildland fire behaviour - i. basic surface interactions, vertical profiles and synoptic structures, Int. J. Wildland Fire 21 (2012) 779–801.

[10] B. E. Potter, Atmospheric interactions with wildland fire behaviour - ii. plume and vortex dynamics, Int. J. Wildland Fire 21 (2012) 802–817.

[11] J. Mandel, J. D. Beezley, A. K. Kochanski, Coupled atmosphere-wildland fire modeling with WRF 3.3 and SFIRE 2011, Geosci. Model. Dev. 4 (2011) 591–610.

[12] J. B. Filippi, X. Pialat, C. B. Clements, Assessment of ForeFire/Meso-NH for wildland fire/atmosphere coupled simulation of the FireFlux experiment, Proc. Combust. Inst. 34 (2013) 2633–2640.

[13] D. Viegas, L. Pita, Fire spread in canyons, Int. J. Wildland Fire 13 (274) (2004) 1–22.

[14] D. Viegas, A. Simeoni, Eruptive behaviour of forest fires, Fire Technol. 47 (2) (2010) 303–320.

[15] M. Nijhuis, Forest fires: Burnt out, Nature 489 (2012) 352–354.

[16] M. Cruz, A. Sullivan, J. Gould, N. Sims, A. Bannister, J. Hollis, R. Hurley, Anatomy of a catastrophic wildfire: The Black Saturday Kilmore East fire in Victoria, Australia, Forest Ecol. Manag. 284 (2012) 269–285.

[17] R. Paugam, M. Wooster, S. Freitas, M. Val Martin, A review of approaches to estimate wildfire plume injection height within large-scale atmospheric chemical transport models, Atmos. Chem. Phys. 16 (2) (2016) 907–925. doi:10.5194/acp-16-907-2016.

[18] E. Koo, P. J. Pagni, D. R. Weise, J. P. Woycheese, Firebrands and spotting ignition in large-scale fires, Int. J. Wildland Fire 19 (7) (2010) 818–843.
URL https://doi.org/10.1071/WF07119

[19] M. Finney, FARSITE: Fire Area Simulator - Model Development and Evaluation, Tech. Rep. February, US Department of Agriculture, Forest Service, Rocky Mountain Research Station (1998).

[20] J.-B. Filippi, F. Bosseur, C. Mari, C. Lac, P. Le Moigne, B. Cuenot, D. Veynante, D. Cariolle, J.-H. Balbi, Coupled atmosphere-wildland fire modelling, J. Adv. Model. Earth Sy. 1 (4) (2009) Quarter 4.

[21] C. Tymstra, R. W. Bryce, B. M. Wotton, S. W. Taylor, O. B. Armitage, Development and structure of Prometheus: the Canadian Wildland Fire Growth Simulation Model, Tech. rep., Natural Resources Canada, Canadian Forest Service, Northern Forestry Centre, Edmonton, Alberta, Information Report NOR-X-417 (2010).

[22] D. Chong, K. G. Tolhurst, T. J. Duff, B. Cirulis, Sensitivity Analysis of PHOENIX RapidFire, Tech. rep., Bushfire CRC, University of Melbourne (2013).

[23] C. Lautenberger, Wildland fire modeling with an eulerian level set method and automated calibration, Fire Safety J. 62, Part C (2013) 289 – 298. doi:http://dx.doi.org/10.1016/j.firesaf.2013.08.014.

[24] A. Sullivan, Wildland surface fire spread modeling, 1990-2007. 2: Empirical and quasi-empirical models, Int. J. Wildland Fire 18 (2009) 369–386.

[25] M. Gollner, A. Trouvé, I. Altintas, J. Block, R. De Callafon, C. Clements, A. Cortes, E. Ellicott, J.-B. Filippi, M. Finney, K. Ide, M.-A. Jenkins, D. Jimenez, C. Lautenberger, J. Mandel, M. C. Rochoux, A. Simeoni, Towards data-driven operational wildfire spread modeling - report of the nsf-funded wifire workshop, Tech. rep., University of Maryland (2015).

[26] M. G. Cruz, M. E. Alexander, Limitations of the accuracy of model predictions of wildland fire behaviour: A state-of-the-knowledge overview, The Forestry Chronicle 89 (3) (2013) 372–383. doi:https://doi.org/10.5558/tfc2013-067.

[27] M. G. Cruz, M. E. Alexander, A. L. Sullivan, J. S. Gould, M. Kilinc, Assessing improvements in models used to operationally predict wildland fire rate of spread, Environmental Modelling & Software 105 (2018) 54 – 63. doi:https://doi.org/10.1016/j.envsoft.2018.03.027.
URL http://www.sciencedirect.com/science/article/pii/S1364815218300161

[28] E. Jimenez, M. Hussaini, S. Goodrick, Quantifying parametric uncertainty in the Rothermel model, Int. J. Wildland Fire 17 (2008) 638–649.

[29] Y. Liu, M. Y. Hussaini, G. Oktenb, Global sensitivity analysis for the rothermel model based on high-dimensional model representation, Canadian Journal of Forest Research 45 (11) (2015) 1474–1479. doi:https://doi.org/10.1139/cjfr-2015-0148.

[30] M. Denham, K. Wendt, G. Bianchini, A. Cortés, T. Margalef, Dynamic data-driven genetic algorithm for forest fire spread prediction, J. Comput. Sci-Neth. 3 (2012) 398–404.

[31] M. C. Rochoux, B. Delmotte, B. Cuenot, S. Ricci, A. Trouvé, Regional-scale simulations of wildland fire spread informed by real-time flame front observations, P. Combust. Inst. 34 (2013) 2641–2647. doi:10.1016/j.proci.2012.06.090.

[32] M. C. Rochoux, S. Ricci, B. Lucor, D.and Cuenot, A. Trouvé, Towards predictive data-driven simulations of wildfire spread - Part 1: Reduced-cost Ensemble Kalman Filter based on a Polynomial Chaos surrogate model for parameter estimation, Nat. Hazards and Earth Syst. Sci. 14 (11) (2014) 2951–2973.

[33] T. Artes, A. Cencerrado, A. Cortes, T. Margalef, D. Rodriguez-Aseretto, T. Petroliagkis, J. San-Miguel-Ayanz, Towards a dynamic data driven wildfire behavior prediction system at european level, Procedia Computer Science 29 (2014) 1216 – 1226. doi:http://dx.doi.org/10.1016/j.procs.2014.05.109.

[34] M. C. Rochoux, C. Emery, S. Ricci, B. Cuenot, A. Trouvé, Towards predictive data-driven simulations of wildfire spread – part ii: Ensemble

kalman filter for the state estimation of a front-tracking simulator of wildfire spread, Nat. Hazards and Earth Syst. Sci. 15 (8) (2015) 1721–1739. doi:10.5194/nhess-15-1721-2015.

[35] C. Zhang, M. C. Rochoux, W. Tang, M. Gollner, J. B. Filippi, A. Trouvé, Evaluation of a data-driven wildland fire spread forecast model with spatially-distributed parameter estimation in simulations of the fireflux i field-scale experiment, Fire Safety J. 91 (2017) 758 – 767, fire Safety Science: Proceedings of the 12th International Symposium. doi:http://dx.doi.org/10.1016/j.firesaf.2017.03.057.

[36] M. C. Rochoux, A. Collin, C. Zhang, A. Trouvé, D. Lucor, P. Moireau, Front shape similarity measure for shape-oriented sensitivity analysis and data assimilation for Eikonal equation, ESAIM: Proceedings and Surveys 63 (2018) 215–236.
URL https://hal.inria.fr/hal-01625575

[37] S. L. Manzello, T. G. Cleary, J. R. Shields, A. Maranghides, W. Mell, J. C. Yang, Experimental investigation of firebrands: Generation and ignition of fuel beds, Fire Safety J. 43 (3) (2008) 226 – 233. doi:https://doi.org/10.1016/j.firesaf.2006.06.010.

[38] N. Sardoy, J. Consalvi, A. Kaiss, A. Fernandez-Pello, B. Porterie, Numerical study of ground-level distribution of firebrands generated by line fires, Combust. Flame 154 (3) (2008) 478 – 488. doi:https://doi.org/10.1016/j.combustflame.2008.05.006.
URL http://www.sciencedirect.com/science/article/pii/S0010218008001600

[39] S. Kortas, P. Mindykowski, J. L. Consalvi, H. Mhiri, B. Porterie, Experimental validation of a numerical model for the transport of firebrands, Fire Safety J. 44 (2009) 1095–1102.

[40] H. A. Perryman, A mathematical model of spot fires and their management implications, Master's thesis, Humboldt State University, Arcata, CA (2009).

[41] H. A. Perryman, C. J. Dugaw, J. M. Varner, D. L. Johnson, A cellular automata model to link surface fires to firebrand lift-off and dispersal, Int. J. Wildland Fire 22 (2013) 428–439.

[42] A. Tohidi, N. Kaye, W. Bridges, Statistical description of firebrand size and shape distribution from coniferous trees for use in metropolis monte carlo simulations of firebrand flight distance, Fire Safety J. 77 (2015) 21 – 35. doi:https://doi.org/10.1016/j.firesaf.2015.07.008. URL http://www.sciencedirect.com/science/article/pii/S0379711215300047

[43] A. Tohidi, Experimental and numerical modeling of wildfire spread via fire spotting, Ph.D. thesis, Clemson University, South Carolina, USA (2016).

[44] A. Tohidi, N. B. Kaye, Stochastic modeling of firebrand shower scenarios, Fire Safety J. 91 (2017) 91 – 102, fire Safety Science: Proceedings of the 12th International Symposium. doi:https://doi.org/10.1016/j.firesaf.2017.04.039.

URL http://www.sciencedirect.com/science/article/pii/S0379711217302461

[45] G. Pagnini, L. Massidda, Modelling turbulence effects in wildland fire propagation by the randomized level-set method, Tech. Rep 2012/PM12a, CRS4, revised Version August 2014. arXiv:1408.6129 (July 2012).

[46] G. Pagnini, Fire spotting effects in wildland fire propagation, in: F. Casas, V. Martínez (Eds.), Advances in Differential Equations and Applications, Vol. 4 of SEMA SIMAI Springer Series, Springer International Publishing Switzerland, 2014, pp. 203–216, iSBN: 978-3-319-06952-4. (eBook: 978-3-319-06953-1).

[47] G. Pagnini, A. Mentrelli, Modelling wildland fire propagation by tracking random fronts, Nat. Hazards Earth Syst. Sci. 14 (2014) 2249–2263.

[48] I. Kaur, A. Mentrelli, F. Bosseur, J.-B. Filippi, G. Pagnini, Turbulence and fire-spotting effects into wild-land fire simulators, Commun. Nonlinear Sci. Numer. Simul. 39 (2016) 300 – 320. doi:http://doi.org/10.1016/j.cnsns.2016.03.003.

[49] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, S. Tarantola, Global Sensitivity Analysis. The Primer, John Wiley & Sons, Ltd, Chichester, UK, 2007. doi:10.1002/9780470725184.

[50] C. Storlie, L. Swiler, J. Helton, C. Sallaberry, Implementation and evaluation of nonparametric regression procedures for sensitivity analysis

of computationally demanding models, Reliab. Eng. Syst. Safe. 94 (11) (2009) 1735–1763. doi:10.1016/j.ress.2009.05.007.

[51] M. Lamboni, H. Monod, D. Makowski, Multivariate sensitivity analysis to measure global contribution of input factors in dynamic models, Reliab. Eng. Syst. Safe. 96 (4) (2011) 450–459. doi:10.1016/j.ress.2010.12.002.

[52] J. J. Ruiz, M. Pulido, T. Miyoshi, Estimating model parameters with ensemble-based data assimilation: A review, Journal of the Meteorological Society of Japan. Ser. II 91 (2) (2013) 79–99. doi:10.2151/jmsj.2013-201.

[53] I. Sobol, Sensitivity analysis for nonlinear mathematical models, Mathematical Modeling and Computational Experiment 1 (4) (1993) 407–414.

[54] B. Sudret, Global sensitivity analysis using polynomial chaos expansions, Reliab. Eng. Syst. Safe. 93 (7) (2008) 964–979. doi:10.1016/j.ress.2007.04.002.

[55] A. Marrel, B. Iooss, B. Laurent, O. Roustant, Calculations of sobol indices for the gaussian process metamodel, Reliab. Eng. Syst. Safe. 94 (3) (2009) 742 – 751. doi:http://dx.doi.org/10.1016/j.ress.2008.07.008.

[56] B. Iooss, A. Saltelli, Introduction to Sensitivity Analysis, in: Handbook of Uncertainty Quantification, Springer International Publishing, 2016, pp. 1–20. doi:10.1007/978-3-319-11259-6_31-1.

[57] L. Le Gratiet, S. Marelli, B. Sudret, Metamodel-Based Sensitivity Analysis: Polynomial Chaos Expansions and Gaussian Processes, in: Handbook of Uncertainty Quantification, Springer International Publishing, 2017, pp. 1–37. doi:10.1007/978-3-319-11259-6_38-1.

[58] N. Owen, P. Challenor, P. P. Menon, S. Bennani, Comparison of surrogate-based uncertainty quantification methods for computationally expensive simulators, SIAM/ASA Journal on Uncertainty Quantification 5 (1) (2017) 403–435. doi:10.1137/15M1046812.

[59] V. Ciriello, V. Di Federico, M. Riva, F. Cadini, J. De Sanctis, E. Zio, A. Guadagnini, Polynomial chaos expansion for global sensitivity analysis applied to a model of radionuclide migration in a randomly heterogeneous aquifer, Stoch. Env. Res. Risk A. 27 (4) (2013) 945–954.

[60] B. Després, G. Poette, D. Lucor, Robust Uncertainty Propagation in Systems of Conservation Laws with the Entropy Closure Method, Springer International Publishing, 2013, pp. 105–149. doi:10.1007/978-3-319-00885-1_3.

[61] S. Dubreuil, M. Berveiller, F. Petitjean, M. Salan, Construction of bootstrap confidence intervals on sensitivity indices computed by polynomial chaos expansion, Reliab. Eng. Syst. Safe. 121 (2014) 263–275. doi:10.1016/j.ress.2013.09.011.

[62] D. Xiu, Numerical Methods for Stochastic Computations: A Spectral Method Approach, Princeton University Press, 2010.

[63] M. De Lozzo, A. Marrel, Sensitivity analysis with dependence and variance-based measures for spatio-temporal numerical simulators, Stoch. Env. Res. Risk A. 31 (6) (2017) 1437–1453.

[64] L. Le Gratiet, C. Cannamela, B. Iooss, A bayesian approach for global sensitivity analysis of (multifidelity) computer codes, SIAM/ASA Journal on Uncertainty Quantification 2 (1) (2014) 336–363. doi:10.1137/130926869.

[65] A. Marrel, G. Perot, C. Mottet, Development of a surrogate model and sensitivity analysis for spatio-temporal numerical simulators, Stoch. Env. Res. Risk A. 29 (3) (2015) 959–974.

[66] J. Oakley, A. O'Hagan, Probabilistic sensitivity analysis of complex models: a bayesian approach, J. Roy. Stat. Soc. B 66 (3) (2004) 751–769. doi:10.1111/j.1467-9868.2004.05304.x.

[67] C. Rasmussen, C. Williams, Gaussian processes for machine learning, MIT Press, 2006.

[68] P. T. Roy, N. El Moçayd, S. Ricci, J.-C. Jouhaud, N. Goutal, M. De Lozzo, M. C. Rochoux, Comparison of polynomial chaos and gaussian process surrogates for uncertainty quantification and correlation estimation of spatially distributed open-channel steady flows, Stoch. Env. Res. Risk A. 32 (6) (2018) 1723–1741. doi:10.1007/s00477-017-1470-4.
URL https://doi.org/10.1007/s00477-017-1470-4

[69] R. Schoebi, B. Sudret, J. Wiart, Polynomial-Chaos-based Kriging, Int. J. Uncertain. Quan. 5 (2) (2015) 171–193.

[70] G. Blatman, Adaptative sparse Polynomial Chaos expansions for uncertainty propagation and sensitivity analysis, Ph.D. thesis, Université Blaise Pascal, Clermont-Ferrand (2009).

[71] G. Blatman, B. Sudret, Adaptative sparse polynomial chaos expansion based on Least Angle Regression, J. Comput. Phys. 230 (6) (2011) 2345–2367.

[72] G. Migliorati, F. Nobile, E. Von Schwerin, R. Tempone, Approximation of quantities of interest in stochastic PDEs by the random Discret L2 Projection on polynomial spaces, SIAM J. Sci Comput. 35 (3) (2013) A1440–A1460.

[73] J. Sethian, Level set methods and fast marching methods, Cambridge University Press, 1999.

[74] S. Osher, R. Fedkiw, Level set methods and dynamic implicit surfaces, Vol. 153, Applied Mathematical Sciences - Springer, 2003.

[75] V. Mallet, D. Keyes, F. Fendell, Modeling wildland fire propagation with level set methods, Comput. Math. Appl. 57 (7) (2009) 1089–1101. doi:10.1016/J.CAMWA.2008.10.089.

[76] C. Lautenberger, Mapping areas at elevated risk of large-scale structure loss using monte carlo simulation and wildland fire modeling, Fire Safety

J. 91 (2017) 768 – 775, fire Safety Science: Proceedings of the 12th International Symposium. doi:https://doi.org/10.1016/j.firesaf.2017.04.014.

[77] A. Mentrelli, G. Pagnini, Front propagation in anomalous diffusive media governed by time-fractional diffusion, J. Comput. Phys. 293 (2015) 427–441.

[78] G. M. Byram, Combustion of forest fuels, in: K. P. Davis (Ed.), Forest Fire: Control and Use, McGraw Hill, New York, 1959, pp. 61–89.

[79] M. E. Alexander, Calculating and interpreting forest fire intensities, Can. J. Bot. 60 (1982) 349–357.

[80] M. Sofiev, T. Ermakova, R. Vankevich, Evaluation of the smoke-injection height from wild-land fires using remote-sensing data, Atmos. Chem. Phys. 12 (4) (2012) 1995–2006. doi:10.5194/acp-12-1995-2012.
URL https://www.atmos-chem-phys.net/12/1995/2012/

[81] D. Xiu, G. Karniadakis, The wiener–askey polynomial chaos for stochastic differential equations, SIAM J. Sci. Comput. 24 (2) (2002) 619–644. doi:10.1137/S1064827501387826.

[82] E. Jones, T. Oliphant, P. Peterson, et al., SciPy: Open source scientific tools for Python, [Online; accessed on 06/02/2019] (2001–).
URL http://www.scipy.org/

[83] K. T. Chu, M. Prodanović, Level set method library (lsmlib), Tech. rep.,

accessed 16th August 2018 (2009).
URL http://ktchu.serendipityresearch.org/software/lsmlib/

[84] C. D. Bevins, Firelib: User manual and technical reference, Tech. rep., US Forest Service, Missoula Fire Sciences Laboratory, Fire Behavior Research Work Unit Systems for Environmental Management, accessed 16th August 2018 (1996).
URL https://www.frames.gov/catalog/935

[85] M. Baudin, A. Dutfoy, B. Iooss, A.-L. Popelin, OpenTURNS: An Industrial Software for Uncertainty Quantification in Simulation, Springer International Publishing, 2017, pp. 2001–2038. doi:10.1007/978-3-319-12385-1-64.
URL https://doi.org/10.1007/978-3-319-12385-1-64

[86] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, Ann. Statist. 32 (2) (2004) 407–499. doi:10.1214/009053604000000067.
URL https://doi.org/10.1214/009053604000000067

[87] D. J. Wales, J. P. K. Doye, Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms, The Journal of Physical Chemistry A 101 (28) (1997) 5111–5116. doi:10.1021/jp970984n.

[88] M. Baudin, K. Boumhaout, T. Delage, B. Iooss, J.-M. Martinez, Numerical stability of Sobol' indices estimation formula, in: 8th International

Conference on Sensitivity Analysis of Model Output,, Réunion Island, 2016.

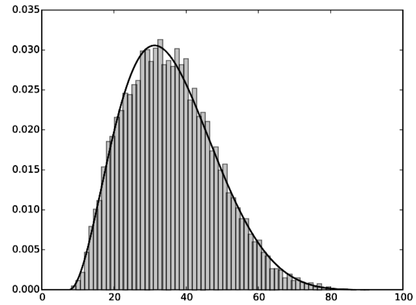
[89] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine Learning in Python, J. Mach. Learn. Res. 12 (2825-2830). arXiv:1201.0490.

[90] P. T. Roy, S. Ricci, R. Dupuis, R. Campet, J.-C. Jouhaud, C. Fournier, Batman: Statistical analysis for expensive computer codes made easy, Journal of Open Source Software 3 (21) (2018) 493. doi:https://doi.org/10.21105/joss.00493.

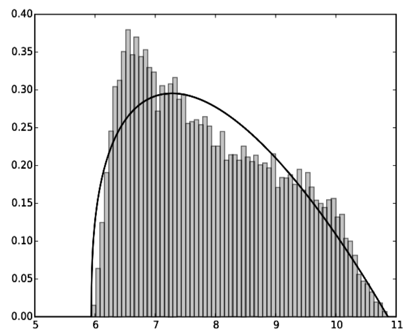
[91] S. W. Taylor, D. G. Woolford, C. B. Dean, D. L. Martell, Wildfire prediction to inform fire management: Statistical science challenges, Stat. Sci. 28 (2013) 586–615.

[92] J. San-Miguel-Ayanz, J. M. Moreno, A. Camia, Analysis of large fires in European Mediterranean landscapes: lessons learned and perspectives, Forest Ecol. Manage. 294 (2013) 11–22.

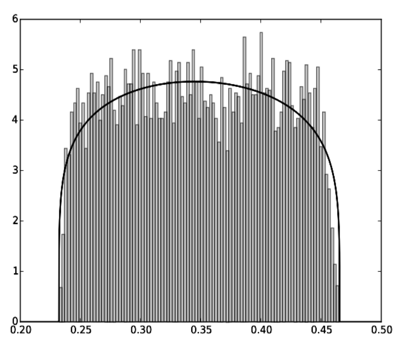
[93] C. Hernandez, C. Keribin, P. Drobinski, S. Turquety, Statistical modelling of wildfire size and intensity: a step toward meteorological forecasting of summer extreme fire risk, Ann. Geophys. 33 (2015) 1495–1506.



(a) Fire-spotting parameter μ.



(b) Fire-spotting parameter σ.



(c) Turbulent diffusion parameter D [$\mathrm{m}^2\,\mathrm{s}^{-1}$].

Figure 1: Histograms and corresponding fits with Beta-distribution (solid lines) for the three parameters μ, σ (fire-spotting effects) and D (turbulence effect) following a Monte Carlo random sampling with 10,000 realizations in the ensemble. Fits performed with the aid of the Python library `SciPy` [82].

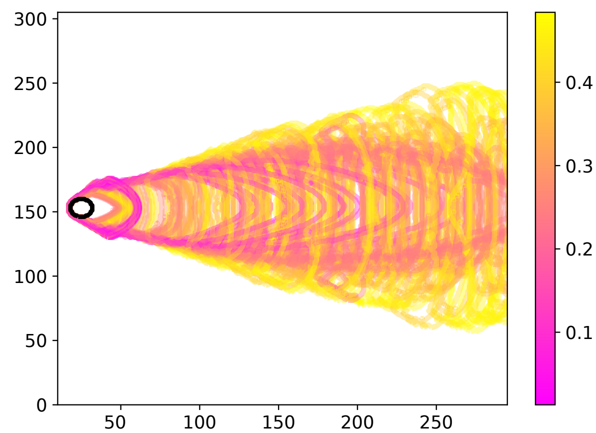


Figure 2: Ensemble of 100 fireline positions over the 2-D computational domain Ω after 50 min of `LSFire+` model integration obtained when varying D, μ and σ as presented in Table 5. The black circle is the initial fireline that is the same for all simulations. The colormap corresponds to the normalized MSR S_t at time $t = 50$ min (Eq. 15).

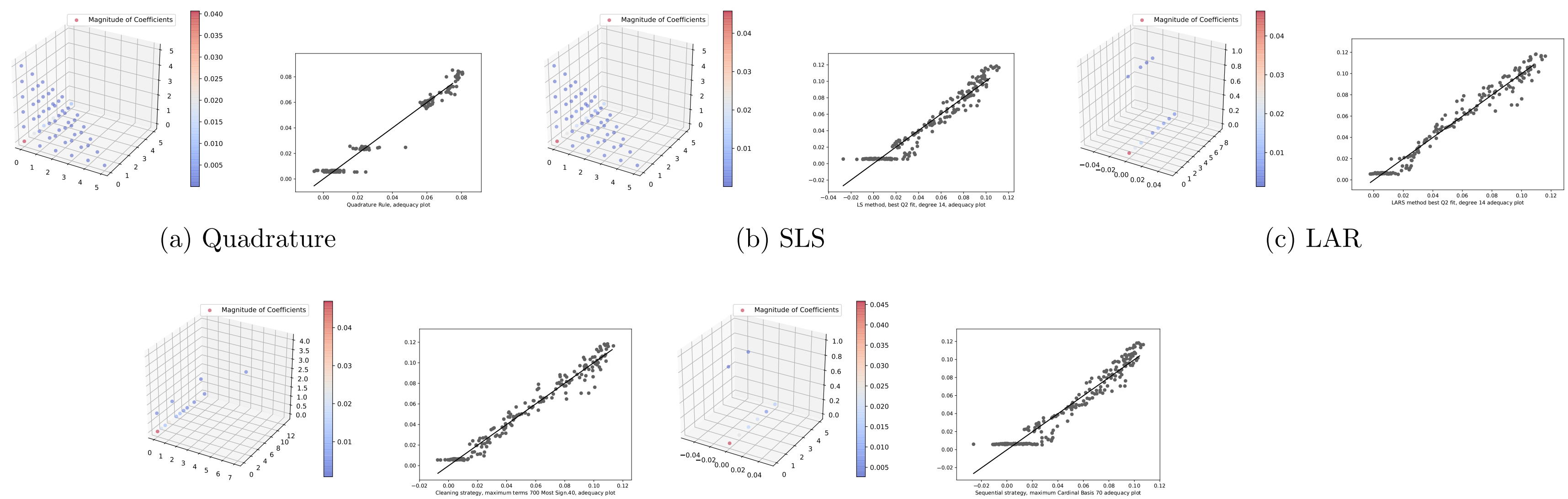
Table 7: Types of surrogate used in this work. Recall that q is the hyperbolic parameter for truncation ($q = 1$ corresponds to linear truncation) and N is the size of the training set.

Name	Truncation	Sparse	Training set
Quad. (Quadrature)	$q = 1, 0.75, 0.5$	No	Gauss quadrature, $N = 216$
SLS (Standard Least-Squares)	$q = 1, 0.75, 0.5$	No	Halton, $N = 216$
LAR (Least-Angle Regression)	$q = 1, 0.75, 0.5$	Yes	Halton, $N = 216$
Cleaning	$q = 1, 0.75, 0.5$	Yes	Halton, $N = 216$
Sequential	$q = 1, 0.75, 0.5$	Yes	Halton, $N = 216$
RBF kernel	–	–	Halton, $N = 216$

Table 8: Error metrics ϵ_{emp} and Q_2 for gPC-expansions and GP-model detailed in Table 7. The size of the training set is $N = 216$. One type of surrogate is built for each of the four observables, A_1, A_2, S_1 and S_2.

	gPC expansion – Linear truncation $(q = 1)$							
	A_1		A_2		S_1		S_2	
	ϵ_{emp}	Q_2	ϵ_{emp}	Q_2	ϵ_{emp}	Q_2	ϵ_{emp}	Q_2
Quad.	$1.4 \cdot 10^{-4}$	0.84	$2.7 \cdot 10^{-4}$	0.86	$5.5 \cdot 10^{-4}$	0.77	$4.6 \cdot 10^{-4}$	0.83
SLS	$3.0 \cdot 10^{-4}$	0.83	$6.3 \cdot 10^{-4}$	0.88	$1.0 \cdot 10^{-3}$	0.74	$2.3 \cdot 10^{-3}$	0.75
LAR	$1.0 \cdot 10^{-4}$	0.99	$4.2 \cdot 10^{-4}$	0.970	$5.0 \cdot 10^{-4}$	0.96	$2.3 \cdot 10^{-3}$	0.95
Cleaning	$1.0 \cdot 10^{-4}$	0.96	$4.1 \cdot 10^{-4}$	0.95	$5.5 \cdot 10^{-4}$	0.96	$1.2 \cdot 10^{-3}$	0.95
Sequential	$3.3 \cdot 10^{-4}$	0.85	$6.7 \cdot 10^{-4}$	0.89	$1.1 \cdot 10^{-3}$	0.77	$2.5 \cdot 10^{-3}$	0.85
	gPC expansion – Hyperbolic truncation $(q = 0.75)$							
	A_1		A_2		S_1		S_2	
	ϵ_{emp}	Q_2	ϵ_{emp}	Q_2	ϵ_{emp}	Q_2	ϵ_{emp}	Q_2
Quad.	$3.7 \cdot 10^{-4}$	0.76	$8.6 \cdot 10^{-4}$	0.77	$1.6 \cdot 10^{-3}$	0.67	$3.7 \cdot 10^{-4}$	0.66
SLS	$1.5 \cdot 10^{-4}$	0.93	$1.8 \cdot 10^{-4}$	0.93	$1.0 \cdot 10^{-3}$	0.84	$2.5 \cdot 10^{-3}$	0.84
LAR	$2.0 \cdot 10^{-4}$	0.94	$5.6 \cdot 10^{-4}$	0.95	$1.0 \cdot 10^{-3}$	0.84	$2.6 \cdot 10^{-3}$	0.86
Cleaning	$9.9 \cdot 10^{-5}$	0.94	$3.3 \cdot 10^{-4}$	0.90	$5.0 \cdot 10^{-4}$	0.96	$1.1 \cdot 10^{-3}$	0.96
Sequential	$1.9 \cdot 10^{-4}$	0.94	$4.7 \cdot 10^{-4}$	0.94	$8.7 \cdot 10^{-4}$	0.86	$1.9 \cdot 10^{-3}$	0.92
	gPC expansion – Hyperbolic truncation $(q = 0.5)$							
	A_1		A_2		S_1		S_2	
	ϵ_{emp}	Q_2	ϵ_{emp}	Q_2	ϵ_{emp}	Q_2	ϵ_{emp}	Q_2
Quad.	$1.8 \cdot 10^{-4}$	0.83	$2.0 \cdot 10^{-4}$	0.87	$6.2 \cdot 10^{-4}$	0.74	$3.6 \cdot 10^{-4}$	0.83
SLS	$1.4 \cdot 10^{-4}$	0.96	$9.6 \cdot 10^{-5}$	0.95	$7.4 \cdot 10^{-4}$	0.86	$1.9 \cdot 10^{-3}$	0.86
LAR	$1.5 \cdot 10^{-4}$	0.97	$4.3 \cdot 10^{-4}$	0.97	$6.5 \cdot 10^{-4}$	0.93	$1.6 \cdot 10^{-3}$	0.94
Cleaning	$8.8 \cdot 10^{-5}$	0.95	$3.3 \cdot 10^{-4}$	0.94	$4.5 \cdot 10^{-4}$	0.92	$9.2 \cdot 10^{-4}$	0.98
Sequential	$1.3 \cdot 10^{-4}$	0.97	$4.2 \cdot 10^{-4}$	0.96	$6.4 \cdot 10^{-4}$	0.93	$1.5 \cdot 10^{-3}$	0.95
	GP model							
RBF	$--$	0.99	$--$	0.98	$--$	0.88	$--$	0.99

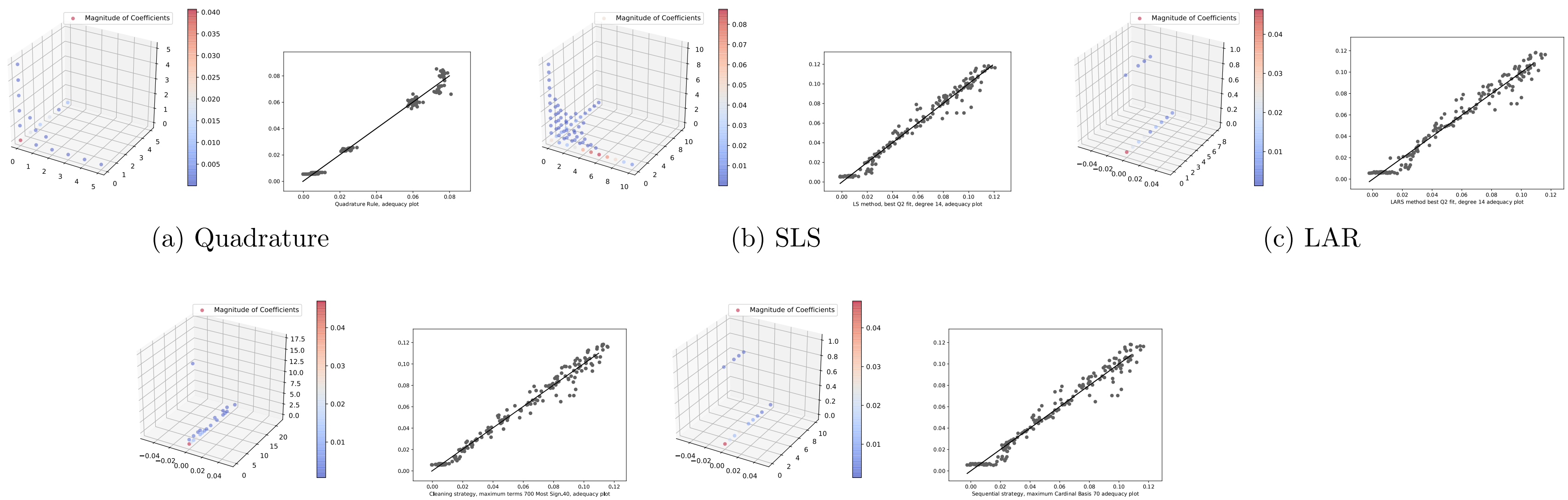
Figure 3: Comparison between quadrature, SLS and sparse (LAR, cleaning, sequential) methods to build the gPC-expansion for the burnt area ratio A_2 using linear truncation. Left: sparsity plots representing the magnitude of the coefficients with respect to the three-dimensional input space ($d = 3$). Right: adequacy scatter plots comparing surrogate (x-axis) and model (y-axis) predictions at the training points. For SLS and LAR, results are obtained with the best fit obtained for varying P.



(a) Quadrature (b) SLS (c) LAR

(d) Cleaning (e) Sequential

Figure 4: Comparison between quadrature, SLS and sparse (LAR, cleaning, sequential) methods to build the gPC-expansion for the burnt area ratio A_2 using linear truncation. Left: sparsity plots representing the magnitude of the coefficients with respect to the three-dimensional input space ($d = 3$). Right: adequacy scatter plots comparing surrogate (x-axis) and model (y-axis) predictions at the training points. For SLS and LAR, results are obtained with the best fit obtained for varying P.



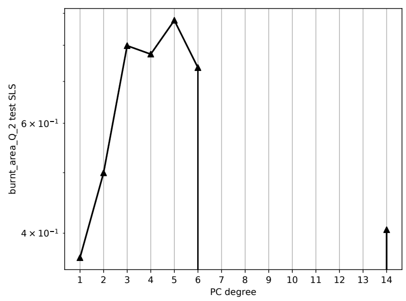
(a) Quadrature

(b) SLS

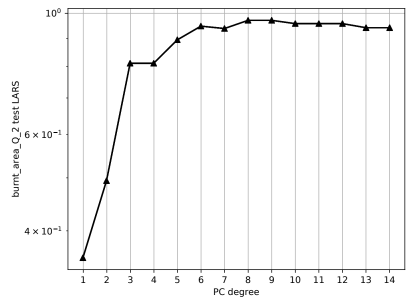
(c) LAR

(d) Cleaning

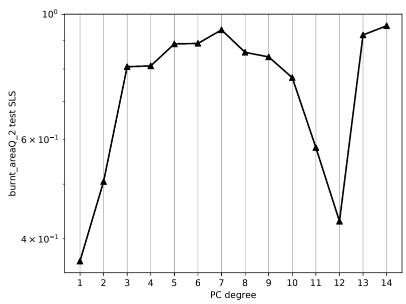
(e) Sequential



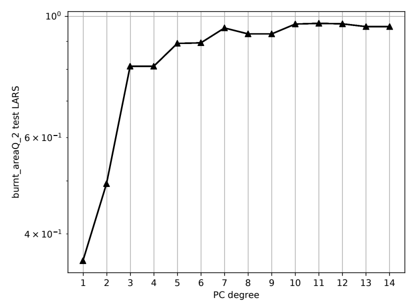
(a) SLS, $q = 1$.



(b) LAR, $q = 1$.



(c) SLS, $q = 0.5$.



(d) LAR, $q = 0.5$.

Figure 5: Sensitivity of the Q_2 predictive coefficient with respect to the total polynomial order P. Comparison of the SLS (a)–(c) and LAR (b)–(d) surrogate methods for linear truncation (top panels) and hyperbolic truncation with $q = 0.5$ (bottom panels) for $1 \leq P \leq 14$.

Table 9: Comparison of Sobol' sensitivity indices associated with the burnt area ratio A_2 and obtained for Halton's low discrepancy sequence.

	S_μ	S_σ	S_D	$S_{T,\mu}$	$S_{T,\sigma}$	$S_{T,D}$
	gPC expansion – Linear truncation $q = 1$					
LAR	0.	0.986	$5.67 \cdot 10^{-3}$	0.	0.994	$1.35 \cdot 10^{-2}$
Cleaning	0.	0.984	$5.89 \cdot 10^{-3}$	$4.70 \cdot 10^{-3}$	0.994	$1.62 \cdot 10^{-2}$
Sequential	0.	0.987	$4.84 \cdot 10^{-3}$	0.	0.995	$1.33 \cdot 10^{-2}$
	GP model					
RBF kernel	$4.59 \cdot 10^{-4}$	0.982	$5.97 \cdot 10^{-3}$	0.001	0.992	0.012

Table 10: Same caption as Table 9 but for the MSR ratio S_2.

	S_μ	S_σ	S_D	$S_{T,\mu}$	$S_{T,\sigma}$	$S_{T,D}$
	gPC expansion – Linear truncation $q = 1$					
LAR	0.	0.948	$1.49 \cdot 10^{-2}$	0.	0.985	$5.22 \cdot 10^{-2}$
Cleaning	0.	0.925	$1.66 \cdot 10^{-2}$	$2.66 \cdot 10^{-3}$	0.983	$7.18 \cdot 10^{-2}$
Sequential	0.	0.954	$1.45 \cdot 10^{-2}$	$7.15 \cdot 10^{-3}$	0.978	$4.63 \cdot 10^{-2}$
	GP model					
RBF kernel	$5.43 \cdot 10^{-4}$	0.941	$9.89 \cdot 10^{-3}$	0.002	0.975	0.047

Table 11: Mean and STD estimate of the burnt area ratio A_2 (left column) and of the MSR ratio S_2 (right column) using linear truncation scheme ($q = 1$), Halton's low discrepancy sequence and gPC or GP surrogate approach.

	A_2	S_2
	gPC expansion – Linear truncation ($q = 1$)	
	mean ± STD	mean ± STD
Quad.	0.0406 ± 0.175	0.102 ± 0.322
SLS	0.0458 ± 0.198	0.114 ± 0.333
LAR	0.0464 ± 0.194	0.114 ± 0.324
Cleaning	0.0469 ± 0.194	0.115 ± 0.327
Sequential	0.0458 ± 0.196	0.113 ± 0.319
	GP model	
	mean ± STD	mean ± STD
RBF kernel	0.0463 ± 0.194	0.114 ± 0.327

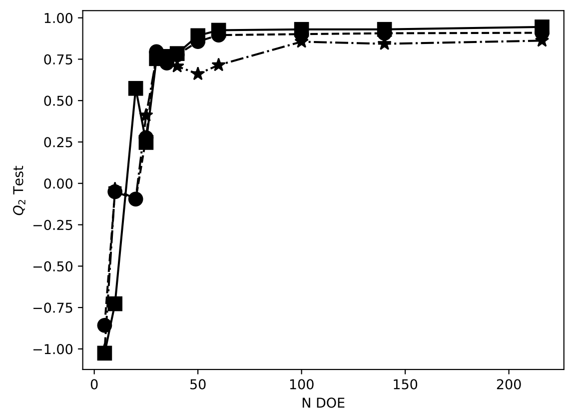


Figure 6: Convergence test with respect to Q_2 predictive coefficient for the LAR gPC-surrogate built using Halton's low discrepancy sequence (cross-validated using the Monte Carlo random sampling). Solid line with square symbols corresponds to linear truncation; dash-dotted line with star symbols corresponds to hyperbolic truncation with $q = 0.75$; and dashed line with circle symbols corresponds to hyperbolic truncation with $q = 0.5$.

Table 12: Sobol' indices (first-order in black and total-order in gray) using LAR gPC-surrogate and linear truncation; $\boldsymbol{\theta} = (U, I, \tau)^T$; $N = 216$. Left: Sobol' indices associated with the burnt area ratio A_2. Right: Sobol' indices associated with the MSR ratio S_2.

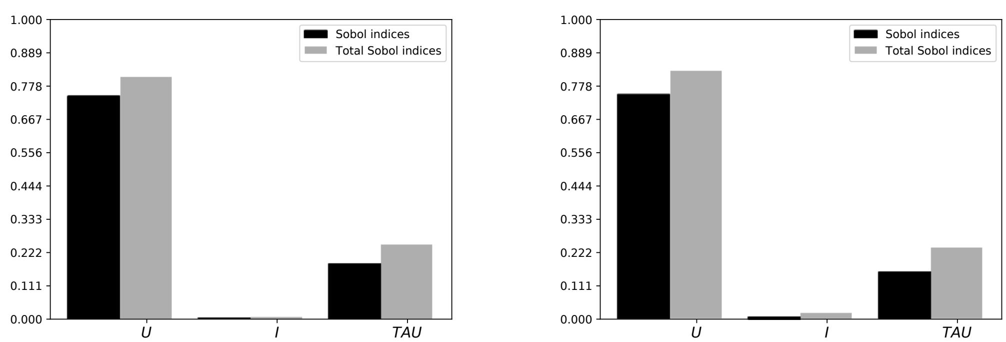
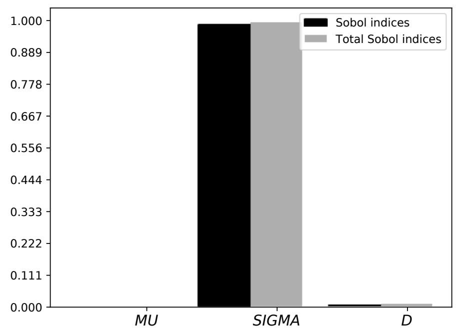


Table 13: Mean and STD of observables A_2 and S_2 as well as error metrics ϵ_{emp} and Q_2 using LAR gPC-surrogate and linear truncation; $\boldsymbol{\theta} = (U, I, \tau)^T$; $N = 216$.

Quantity of interest	Mean	STD	ϵ_{emp}	Q_2
A_2	0.07	0.06	$9 \cdot 10^{-4}$	0.95
S_2	0.19	0.13	$2 \cdot 10^{-3}$	0.96

Table 14: Same caption as in Table 12 but for $\boldsymbol{\theta} = (\mu, \sigma, D)^T$.



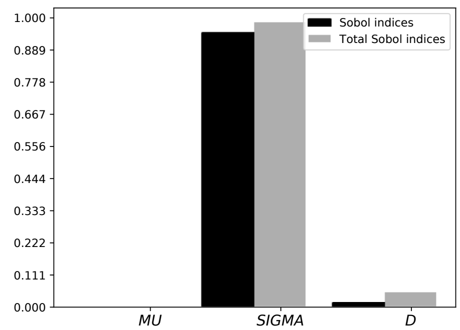


Table 15: Same caption as in Table 13 but for $\boldsymbol{\theta} = (\mu, \sigma, D)^T$.

Quantity of interest	Mean	STD	ϵ_{emp}	Q_2
A_2	0.05	0.04	$4 \cdot 10^{-4}$	0.97
S_2	0.11	0.11	$2 \cdot 10^{-3}$	0.95