Multioutput Feature Selection for Emulation and Sensitivity Analysis

Jorge Vicent, Luca Martino, Jochem Verrelst, Juan Pablo Rivera Caicedo and Gustau Camps-Valls *Fellow, IEEE*.

Abstract

Statistical regression methods are widely used in remote sensing applications but tend to lack physical interpretability. In this paper, we introduce a methodological framework to improve model emulation and its understanding with machine learning feature selection. Our wrapper-forward feature selection method seamlessly integrates physics knowledge into model emulation, improving the trade-off between accuracy and interpretability. We illustrate our methodology by applying it to atmospheric radiative transfer models in the context of global sensitivity analysis (GSA) and emulation. Our approach consistently aligns with variance-based GSA, pinpointing the critical features of aerosol properties, solar zenith angle, and water vapor. While our physically-based emulators yield only a modest accuracy improvement of 0.2% over conventional Gaussian Processes emulators, its introduction signifies a step forward to physics-aware machine learning-based emulation. The emulator performance remains steadfast, unaffected by substantial changes, further underscoring the reliability of our approach.

Index Terms

Emulation, feature selection, Gaussian processes, hyperspectral, physics-aware machine learning

I. Introduction

Statistical regression methods are widely used in remote sensing applications, such as classification, biophysical parameters retrieval, and emulation [1], [2]. These methods offer numerous

Manuscript submitted October 30, 2023. Luca Martino was supported by the Young Researchers R&D Project, ref. Num. F861 (AUTO-BA- GRAPH) funded by the Community of Madrid and Rey Juan Carlos University and by Agencia Estatal de Investigacion AEI (project SP-GRAPH, ref. num. PID2019-105032GB-I00). Jochem Verrelst was supported by the European Research Council (ERC) under the ERC-2017-STG SENTIFLEX project (grant agreement 755617). Gustau Camps-Valls was supported by the ERC under the ERC-CoG-2019 USMILE project (grant agreement 855187) and the Marie Curie ETN iMIRACLI (860100). J. Vicent, J. Verrelst and G. Camps-Valls are with Image Processing Laboratory, University of Valencia, Spain. E-mail: [jorge.vicent, jochem.verrelst, gustau.camps]@uv.es. L. Martino is with the Universitá di Catania, Italia. Email: luca.martino@unict.it. J.P. Rivera is with CONAHCYT-UAN, Universidad Autónoma de Nayarit, Mexico. Email: jprivera@uan.edu.mx

advantages, including accuracy, adaptability, and computational efficiency. Nevertheless, the mathematical implementation and hyperparameters of the underlying machine learning algorithm tend to lack physical interpretability. This opacity hampers our ability to understand how the predictions are generated. Physics-aware machine learning is an approach that addresses this issue by incorporating physics knowledge into machine learning models [3]–[6]. This has the potential to enhance the accuracy, reliability, and performance of statistical regression models while providing a degree of model explainability that helps us to better understand the relationships between the input and output variables. Incorporating physical knowledge into a statistical regression method can be done in various ways, such as using physics-based features [7], implementing physical constraints into the model [8], or generating training datasets with physical models [1], [9]. Supervised feature selection falls within the first category as a method to construct physics-aware statistical regression models. Feature selection is a technique that aims at reducing the number of input variables in a model by selecting the most relevant ones based on their impact on the model outputs [10], [11]. There are three main mechanisms for feature selection: (1) wrapper, (2) filter, and (3) intrinsic (or embedded) methods. The most important sub-class of the wrapper methods is called stepwise methods: they create multiple regression models varying the subset of selected features and choosing the most accurate one [10]. Filter methods use ad hoc statistical techniques to assess the relationship between input and output variables [12]. Intrinsic methods refer to statistical regression algorithms that automatically perform feature ranking and selection during model training. Examples of intrinsic methods are automatic relevance determination (ARD) applied to Gaussian Processes (GP) [13], [14], neural networks [15] and random forests [16]. All these approaches have been widely used in remote sensing applications [17]–[19].

The main objective of this study is to enhance the physical awareness and explainability of emulators by applying a feature selection algorithm. Emulation is a method to approximate the outputs of a deterministic model using statistical regression algorithms such as e.g., GP, neural networks, or random forests [20], [21]. The key advantage of emulators is that they provide high-accuracy predictions and fast runtimes [22], [23]. We use a wrapper feature selection method in which relevant features are sequentially added to a regression model to minimize a cost function. This iterative process creates a ranking of features from the most to the least relevant. Unlike the filter-based methods, the wrapper-based approach prioritizes the minimization of a cost function and thus the optimization of an emulator accuracy while considering the impact of features on model performance. The selection of optimal features is then determined using the *spectral information criterion* (SIC) [24]. SIC unifies several information criterion approaches [25] such as the Bayesian or Akaike information criteria [26], [27], containing them as special cases (see Tab. I). We apply this feature selection method to spectral (multioutput) data generated by atmospheric radiative transfer models (RTM). Specifically, we first derive a global sensitivity analysis (GSA) directly from the feature selection method in §II. GSA quantifies how input variables affect a model's output, considering all possible values and interactions [28]. GSA allows us to identify the most relevant input features of an RTM based on their influence on the output spectral data. This information can be used to improve the accuracy of RTM-based applications, reduce uncertainty, and make better decisions [29]. We focus on emulating atmospheric radiative transfer models in §III. We then evaluate and discuss the accuracy of the proposed feature selection method for emulation and GSA in §IV and §V. The paper ends with a final discussion in Section VI.

II. Theoretical description

A. Mathematical nomenclature

Let us consider a dataset $\{\mathbf{x}_i, y_i\}_{i=1}^n$, where $\mathbf{x}_i = [x_{i1}, \ldots, x_{id}]^\top \in \mathbb{R}^d$ and $y_i = g(\mathbf{x}_i) \in \mathbb{R}$ represent the inputs and outputs of a deterministic model $g(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$. We define the *partial* model $g_k(\mathbf{z}^{(k)}) : \mathbb{R}^k \to \mathbb{R}$ as a reduced version of $g(\mathbf{x})$ that uses a subset of $k \leq d$ features so that $g_k(\mathbf{z}_i^{(k)})$ is as close as possible to $g(\mathbf{x}_i)$. Here, $\mathbf{z}_i^{(k)} = [z_{i1}, \ldots, z_{ik}]^\top \in \mathbb{R}^k$ with each $z_{ij} \in \{x_{i1}, \ldots, x_{id}\}$ and so that $j = 1, \ldots, d$ cannot be repeated in $\mathbf{z}_i^{(k)}$. We consider that the order of the features inside g_k does not matter e.g., $g_k(z_{i1}, z_{i2}, \ldots, z_{ik}) = g_k(z_{i3}, z_{ik}, \ldots, z_{i2})$. For constructing $\mathbf{z}_i$ we have to choose k features over the d possible features without repetition (order does not matter). Namely, we have C_k^d different combinations given in (1) for building $\mathbf{z}_i^{(k)}$.

$$C_k^d \equiv \begin{pmatrix} d \\ k \end{pmatrix} = \frac{d!}{k!(d-k)!}. \tag{1}$$

In the extreme case where $k = d$ we just have one possible choice, $\mathbf{z}_i^{(d)} = \mathbf{x}_i$, and thus $g_d(\mathbf{z}_i^{(d)}) \equiv g(\mathbf{x}_i) = y_i$.

B. Ranking of features

We aim to rank the d features of the input space taking into account their impact on the output of $g(\mathbf{x})$ by using a *forward selection* method [30], [31]. This method consists in adding recursively variables minimizing a cost function χ_k that measures the difference between $g(\mathbf{x})$ and $g_k(\mathbf{z}^{(k)})$. The method starts searching for the most significant single variable model (in terms of the cost function value), i.e., considering a *partial* model with only one feature (k=1). This search is repeated considering a *partial* model with k=2 variables, re-estimating the model for each pair of variables, including and keeping the previously selected variable. We iterate the procedure until reaching a complete model of $k = d$ variables. This procedure provides a sequence of variables that will be the final ranking. Since the forward selection method is based on minimizing the cost function χ_k, the produced ranking depends directly on the model $g(\mathbf{x})$. Namely, the *importance* associated with each feature is related to the output that we are analyzing. Common choices for χ_k are the L_p norms, defined by Eq. (2), and their relative counterparts:

$$\chi_k = \left[\frac{1}{n}\sum_{i=1}^{n}\left|y_i - g_k(\mathbf{z}_i^{(k)})\right|^p\right]^{1/p}. \tag{2}$$

Note that the produced ranking depends on the implemented cost function although, generally, the ranking positions of the most important features remain usually unaltered by a slight change in the cost function (e.g., changing L_1 by L_2 norms). The forward selection method is given in pseudo-code 1.

Algorithm 1 Forward selection

```
Input: The dataset $\{\mathbf{x}_i, y_i\}_{i=1}^n$, an empty vector $\mathbf{z}_i^{(0)}$, the set of indices $\mathcal{J}_0 = \{1, 2, \dots, d\}$, and
  $V(0) = \chi_0 \equiv \left[\frac{1}{n}\sum_{i=1}^n |y_i - \bar{y}|^p\right]^{1/p}$ where $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$
  for $k$=1 to $d$ do
    for $j \in \mathcal{J}_{k-1}$ do
      1. Set $\widetilde{\mathbf{z}}_{ij}^{(k)} = [\mathbf{z}_i^{(k-1)}, x_{ij}]$
      2. Compute $\chi_{kj} \equiv \left[\frac{1}{n}\sum_{i=1}^n |y_i - g_k(\widetilde{\mathbf{z}}_{ij}^{(k)})|^p\right]^{1/p}$
    end for
    3. Set $j^* = \arg\min_j \chi_{kj}$ and $V(k) = \min_j \chi_{kj}$.
    4. Set $\mathbf{z}_i^{(k)} = [\mathbf{z}_i^{(k-1)}, x_{ij^*}]$.
    5. Remove $j^*$ from $\mathcal{J}_{k-1}$ and set $\mathcal{J}_k = \mathcal{J}_{k-1}\backslash\{j^*\}$.
  end for
Output: $\mathbf{z}_i^{(d)}$ and $V(k)$ for $k = 0, \dots, d$
```

Note that the number of indices in $\mathcal{J}_k$ is decreasing with each iteration of k, starting from d indices at k=0 and finishing with one index when k=d=1.

As presented in the pseudo-code 1, the forward selection method has two main drawbacks. The first drawback is related to the definition of the *partial* model $g_k(\mathbf{z}^k)$ as a reduced version of $g(\mathbf{x})$. The *partial* model $g_k(\mathbf{z}^k)$ takes values in a smaller space than $g(\mathbf{x})$ since $\mathbf{z}^k$ has a smaller dimension than $\mathbf{x}$ (except for the last iteration, where they have the same length). Therefore, in order to obtain a *partial* model, we should integrate all the features that are not included in $\mathbf{z}^k$. As a consequence, the partial models are analytically unknown. The second drawback is that the model $g(\mathbf{x})$ is computationally slow in real-life scenarios (e.g., atmospheric RTMs [32]). Thus, the forward selection method would be impractical given the many simulations needed. To overcome these two drawbacks, we use instead a regression function $\widehat{g}_k(\mathbf{z}^{(k)}) : \mathbb{R}^k \to \mathbb{R}$ that approximates $g(\mathbf{x})$ with a much faster run time. Given a specific choice of the elements z_{ij} in $\mathbf{z}^{(k)}$, $\widehat{g}_k(\mathbf{z}_i^{(k)})$ is obtained from the regression $\mathbf{z}_i^{(k)} \to y_i$, thus linking a subset of all possible features to the output y. Possible choices for regression functions range from simple parametric models to statistical regression methods called *emulators* [33]–[35]. The specific regression method employed has an impact on the ranking of features given that the cost function depends directly on $\widehat{g}_k(\mathbf{z}_i^{(k)})$. However, if the used regression method is accurate enough and the parameters (or hyper-parameters) well-tuned, the obtained ranking should not change substantially for a change of the regression function. Although emulators are a priori best suited due to their higher flexibility and accuracy, they need to be re-trained every time that a new feature is added. This makes them slow for the forward selection method. Instead, a well-designed parametric model can capture the main dependencies of $g(\mathbf{x})$ while being fast to “train” (e.g., through least-squares fitting) and run. In the application discussed in this work (i.e., atmospheric RTMs), a d-dimensional 2nd degree polynomial fitting is a pragmatic solution given the smooth dependencies of the output spectral data (e.g., transmittance) to the input atmospheric and geometric features.

So far we have presented the case of a single output model $g(\mathbf{x})$. For a multi-output model $\mathbf{g}(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}^b$, the forward selection method can still be applied with a few considerations. First, since the dataset $\{\mathbf{x}_i, \mathbf{g}(\mathbf{x}_i)\}_{i=1}^n$ is now multi-output, the regression function must also be multi-output, $\widehat{\mathbf{g}}_k(\mathbf{z}^{(k)}) : \mathbb{R}^k \to \mathbb{R}^b$. Second, since the forward selection method relies on a scalar cost function χ_k, Eq. (2) must be adapted. A straightforward option is to perform an average of the multiple outputs:

$$\chi_k = \frac{1}{b} \sum_{\lambda=1}^{b} \left[\frac{1}{n} \sum_{i=1}^{n} \left| g^{\lambda}(\mathbf{x}_i) - g_k^{\lambda}(\mathbf{z}_i^{(k)}) \right|^p \right]^{1/p}, \tag{3}$$

where the superscript $\lambda = 1, \dots, b$ identifies a specific dimension in the output data e.g., $\mathbf{g}(\mathbf{x}_i) = [y_i^1, \dots, y_i^{\lambda}, \dots, y_i^b]$.

C. Selection of the number of relevant features

The minimum error function $V(k)$, for $k = 0, \dots, d$, is generally a decreasing function. Several procedures (also denoted as *stopping rules*) can be applied for the selection of the optimal number of relevant features. Clearly, a naive method is to set a threshold value (or a percentage) for the prediction error, or by visual inspection of the error curve, i.e., the so-called *elbow method* [36], [37]. In the classical variables selection analysis in multiple linear regression, statistical tests (e.g., F-test, t-test and Wald test) are employed sequentially to decide whether individual variables should be included in the model, and a stopping rule based on p-values under the null-hypothesis that the coefficients in the multiple linear regression are zero [38], [39]. Namely, significant variables (with a corresponding coefficient with smaller p-values) get included, but insignificant ones do not (with a corresponding coefficient with higher p-values). Other approaches rely on the so-called *information criteria* [25], [40]. Here, we consider the recently proposed SIC [24]. Similarly to other information criteria, a cost function is constructed,

$$C(k, \gamma) = \underbrace{V(k)}_{\text{fitting}} + \underbrace{\gamma k}_{\text{penalization}}, \qquad k = 0, \dots, d, \tag{4}$$

where γ >0. The first term is a fitting term, whereas the second term is a linear penalty for the model complexity (see Fig. 1).

In a standard information criterion, γ is a chosen constant that specifies the criterion, and the first term is based on a maximum likelihood value $\ell_{\texttt{max}}$, specifically, $V(k) = -2 \log \ell_{\texttt{max}}$. Assuming this choice of $V(k)$, the expression (4) encompasses several well-known information criteria proposed in the literature (see Tab. I), which differ for the choice of γ [25], [40].

Spectral[1] Information Criterion (SIC). In SIC [24], the fitting term $V(k)$ can be any non-increasing function, e.g., the minimum error obtained in the forward selection ranking. Moreover,

[1] In this context, the term *spectral* refers to the extraction of geometric information from the curve $V(k)$. To each γ, the SIC method associates a positive coefficient (as the module transformation in a Fourier analysis of a signal), obtaining a *spectra* as a function of γ. It is important not to confuse this term with the electromagnetic *spectral* data simulated by an RTM.

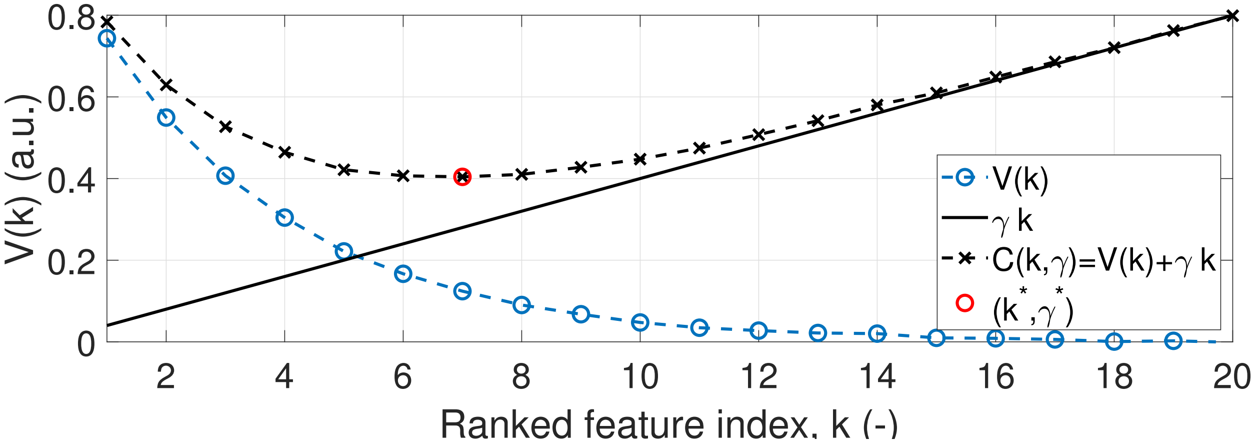

Fig. 1. Example of function $V(k)$, a penalization term γk, and the corresponding cost function $C(k, \gamma)$ (shown with dots).

TABLE I

SPECIAL CASES OF INFORMATION CRITERIA CONTAINED IN SIC, WITH THE PROPER CHOICE OF γ. NOTE THAT n DENOTED THE NUMBER OF DATA POINTS AND $\ell_{\max}$ IS THE MAXIMUM VALUE REACHED BY A LIKELIHOOD FUNCTION (REPRESENTING THE SPECIFIC APPLICATION).

Information criterion	**Choice of γ**	$V(k)$
Bayesian-Schwarz [26]	$\log n$	$-2 \log \ell_{\max}$
Akaike [27]	2	$-2 \log \ell_{\max}$
Hannan-Quinn [41]	$\log(\log(n))$	$-2 \log \ell_{\max}$
Automatic Elbow Detector [42]	$\frac{V(0)}{\min[\arg\min V(k)]}$	any
SIC [24]	all	any

SIC obtains the distribution of minima of the cost function $C(k, \gamma)$ as γ varies in the interval $[0, \gamma_{\mathtt{max}}]$, where $\gamma_{\mathtt{max}}$ is defined as $\gamma_{\mathtt{max}} = \{\min \gamma : \ \arg\min_k C(k, \gamma) = 0\}$, i.e., is the minimum value of γ which provides the strongest possible model penalization (choosing a model with zero variables all the features are irrelevant for predicting the output y). The value of $\gamma_{\mathtt{max}}$ can be analytically obtained as

$$\gamma_{\mathtt{max}} = \max_k \left[\frac{V(0) - V(k)}{k} \right], \quad \text{for } k = 1, \dots, d. \tag{5}$$

Since above we consider $k = 1, \dots, d$, we can perform an exhaustive search and obtain $\gamma_{\mathtt{max}}$ from Eq. (5). The SIC approach is inspired by the idea of "integrating out" γ, i.e., to remove the dependence of γ and hence avoid picking a specific value of γ. To each value of k', SIC associates an interval of γ values, $\mathcal{S}_{k'} \subset [0, \gamma_{\mathtt{max}}]$, such that for each $\gamma^* \in \mathcal{S}_{k'}$, then $\arg\min_k C(k, \gamma^*) = k'$.

These intervals, for $k = 1, \dots, d$, form a partition of $[0, \gamma_{\texttt{max}}]$, i.e.,

$$\mathcal{S}_1 \cup \mathcal{S}_2 \dots \cup \mathcal{S}_d = [0, \gamma_{\texttt{max}}], \tag{6}$$

and $\mathcal{S}_k \cap \mathcal{S}_j = 0$, for all $k \neq j$. Observe that, by construction, $\mathcal{S}_0 = \emptyset$ due to the definition of $\gamma_{\texttt{max}}$. Fig. 2 provides a graphical representation. Furthermore, we can use the information provided by the measures $|\mathcal{S}_k|$, defining the weights $\bar{w}_k \propto |\mathcal{S}_k|$, i.e.,

$$\bar{w}_k = \frac{|\mathcal{S}_k|}{\sum_{j=0}^{d} |\mathcal{S}_j|} = \frac{|\mathcal{S}_k|}{\sum_{j=1}^{d} |\mathcal{S}_j|}, \tag{7}$$

where $|\mathcal{S}_0| = 0$. Note that $\bar{w}_k$, for $k = 1, \dots, d$, defines a probability mass function, $\sum_{k=1}^{d} \bar{w}_k = 1$. These weights can be computed by the Monte Carlo method [24].

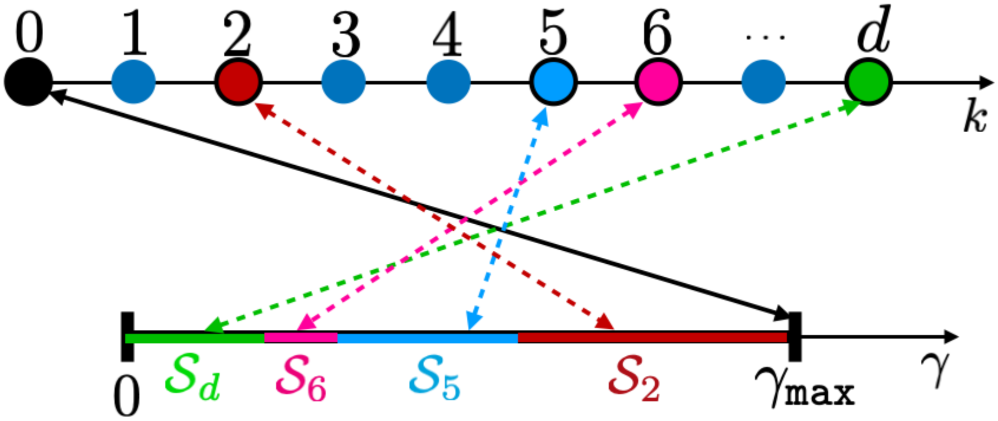

Fig. 2. Graphical representation of intervals $\mathcal{S}_k$ and measures $|\mathcal{S}_k|$ for all k.

A first output of SIC is the set $\mathcal{E}$ of indices such that the corresponding weights are non-zero:

$$\mathcal{E} = \{\text{all } k : \ \bar{w}_k > 0\} = \{k_E^{(1)}, k_E^{(2)}, \dots, k_E^{(J)}\}. \tag{8}$$

They can be interpreted as a possible "elbow" of the curve $V(k)$, i.e., any possible selected model is represented by one index $k_E^{(j)}$. We have denoted $J = |\mathcal{E}|$ with $J \leq d$ and, in some cases, $J \ll d$. This is because some value $k' \neq 0$ could never be a minimum so that $|\mathcal{S}_{k'}| = 0$. Therefore, we can have a sensible reduction of the number of possible models to choose from. In order to select just one model, the more conservative solution is $k_E = \max k_E^{(j)}$ choosing the more complex model. However, the suggestion in [24] is to define the cumulative sum of the first m weights i.e., $W_m = \sum_{i=1}^{m} \bar{w}_i$, with $1 < m \leq d$ and choose as "elbow" the index defined as

$$k_E = \min\{k :\ W_k \geq \ell\}, \qquad \text{with} \qquad \ell \geq 0.9, \tag{9}$$

where ℓ is a confidence level. A more conservative choice of ℓ=0.98 selects a more complex model within $\mathcal{E}$.

D. Global Sensitivity Analysis (GSA) and Emulation

In the context of **GSA**, the forward selection method provides a direct way to compute sensitivity indices (SI) using the error magnitude $V(k)$. We define the SI (ranging from 0% to 100%) for each ranked feature in $\mathbf{z}^{(d)}$ as follows:

$$SI(k) = \frac{100 \cdot [V(k) - V(k-1)]}{\sum_{i=1}^{d}[V(i) - V(i-1)]}. \tag{10}$$

Indeed, we expect that the decrease in the error magnitude $V(k)$ will be the highest in the first feature of the ranking (k=1) and the lowest in the last feature (k=d). For multi-output models, applying the forward selection method as a regression-based GSA algorithm is reduced to using the single-output version and looping over each output dimension.

In the context of **emulation**, feature selection can be applied to include only the relevant input features when training and running a statistical regression algorithm, thus making the emulator model more accurate. Without loss of generality, in this work, we use GP emulators (see [35] for a complete description) and we propose two options to use feature selection. The first option is to apply feature selection directly on the multi-output data so that, instead of $\mathbf{x}_i \rightarrow \mathbf{g}(\mathbf{x}_i)$, the emulator does now the regression $\mathbf{z}_i^{(k)} \rightarrow \mathbf{g}(\mathbf{x}_i)$. In this option, the selected features $\mathbf{z}_i^{(k)}$ remain the same for all the individual GPs applied in each PCA component [2]. In practice, this option is equivalent to a GP emulator with a Gaussian kernel (see Eq. (11)) with only two hyper-parameters (θ_f and θ_l) but with a reduced number of input dimensions.

$$k(\mathbf{x}, \mathbf{x}^*) = \theta_f^2 \exp\left(-\frac{(\mathbf{x}-\mathbf{x}^*)^\top(\mathbf{x}-\mathbf{x}^*)}{2\theta_l^2}\right). \tag{11}$$

The second option considers that each PCA component has a different sensitivity to each feature. Hence, we use the sensitivity index, $SI(k)$, to adjust the influence of features within

[2] A multi-output GP emulator is achieved by reducing the dimensionality of the outputs by PCA and training an individual GP for each component [35].

the regression model. This is achieved by defining the Mahalanobis Gaussian kernel as:

$$k(\mathbf{x},\mathbf{x}^*) = \theta_f^2 \exp\left(-\frac{(\mathbf{x}-\mathbf{x}^*)^\top W(\mathbf{x}-\mathbf{x}^*)}{2\theta_l^2}\right), \tag{12}$$

where W is a $d \times d$ diagonal matrix with values $W_{kk} = SI(k)^2$ for k=1 to d. Eq. (12) can be re-written as:

$$k(\mathbf{x},\mathbf{x}^*) = \theta_f^2 \exp\left(-\sum_{k=1}^{d} \frac{W_{kk}(x_k - x_k^*)^2}{2\theta_l^2}\right), \tag{13}$$

which is equivalent to the ARD-Gaussian kernel in Eq. (14):

$$k(\mathbf{x},\mathbf{x}^*) = \theta_f^2 \exp\left(-\sum_{k=1}^{d} \frac{(x_k - x_k^*)^2}{2\theta_{l,k}^2}\right), \tag{14}$$

where $\theta_{l,k}^2 \equiv \theta_l^2/W_{kk}$. The distinction between the ARD and Mahalanobis kernels lies in the fact that the ARD version utilizes one scale-length hyper-parameter for each input dimension, whereas the Mahalanobis kernel employs a single hyper-parameter. Both kernels provide a degree of model explainability, wherein less relevant features are characterized by higher values of their associated scale-lengths, resulting in reduced impact on the model.

III. Materials and Methods

A. *Simulated dataset and tools*

To test the performance of the proposed feature selection method, we used MODTRAN6 to create a database of atmospheric transfer functions: path radiance ($\mathbf{L}_0$), direct/diffuse at-surface solar irradiance ($\mathbf{E}_{dir/dif}$), direct/diffuse surface-to-satellite transmittance ($\mathbf{T}_{dir/dif}$), and spherical albedo ($\mathbf{S}$). Each of these transfer functions represents the model $\mathbf{g}(\mathbf{x})$ described in the previous section. The simulations were performed following our previous work [34], [35]. We used a Latin Hypercube Sampling (LHS) [43] with n=500 data points in the input parameter space (d=9) with the boundaries given in Table II. MODTRAN6 was configured to cover the spectral range 400-2500 nm with a sampling of 5 cm^{-1} (i.e., 0.08 nm to 3 nm, b=4200). To assess the impact of feature selection on the accuracy of an emulator, we also simulated a *reference* dataset of m=10000 samples generated with LHS distribution and the same input variables and ranges as in Table II.

Both datasets were generated with the Atmospheric Look-up table Generator (ALG) toolbox v3.2 [35], [44] and are accessible from https://doi.org/10.5281/zenodo.7826005. ALG is a soft-

TABLE II
RANGE OF RTM INPUT VARIABLES. VIEWING ZENITH IS SET TO 0°. SENSOR ALTITUDE IS AT SATELLITE LEVEL. COLORED SQUARE IDENTIFIES EACH FEATURE IN SECTION IV.

Input variables	**Units**	**Min.**	**Max.**
■ O_3 column concentration (O3)	[amt-cm]	0.25	0.45
■ Columnar Water Vapour (CWV)	[g·cm^{-2}]	0.2	4
■ Aerosol Optical Thickness (AOT)	unit-less	0.04	0.6
■ Asymmetry parameter (g)	unit-less	0.5	0.85
■ Ångström exponent (α)	unit-less	0.1	2
■ Single Scattering Albedo (SSA)	unit-less	0.8	1
■ Surface elevation (h)	[km]	0	3
■ Solar Zenith Angle (SZA)	[deg]	0	70
■ Relative Azimuth Angle (RAA)	[deg]	0	180

ware tool that facilitates users to configure and run simulations from a suite of atmospheric RTMs. ALG includes all the algorithms described in Section II to automate feature selection and use it in the context of GSA and emulation. The ALG tool can be downloaded from www.artmotoolbox.com. The emulator function, which includes the feature selection and GSA algorithms, is accessible from https://github.com/jorviser/AlgEmulator for standalone use.

B. Assessment methodology

First, we studied the behavior of the feature selection method by assessing the impact of the cost function (L_1, L_2, and their relative counterparts) on the ranking of features. The analysis was done for each transfer function individually on 4 specific wavelengths (i.e., a single output model) that have different sensitivity to the input variables: 400 nm (aerosols), 603 nm (ozone), 761 nm (surface elevation), and 940 nm (water vapor). To illustrate the behavior of the forward selection method, we plot the curve $V(k)$ for the selected cost functions and display the number of selected features.

Second, in the context of GSA, we calculate the SI for the 4 selected wavelengths and all six atmospheric transfer functions. The GSA results using the presented forward selection method are compared with two independent methods. The first one is the variance-based GSA method proposed by Satelli et al. [12], for which we exploit the total SI (i.e., taking into account covariance in the data). The second one is based on the length-scale hyperparameters of an ARD-Gausssian GP regression using Matlab's `fitrgp` function [14]. Although the ARD-kernel is not per se a GSA method, its scale-length hyperparameters give information about the relevance of the input features. Here, we define the SI as the normalized inverse of the scale-lengths in Eq. (15):

$$SI_{ARD,k} = \frac{100\theta_{l,k}^{-1}}{\sum_{k=1}^{d}\theta_{l,k}^{-1}}. \tag{15}$$

The GSA analysis is then expanded to the entire spectral range, fixing the cost function to a L_1 norm and only applying the presented method and the variance-based method.

Third, we assessed the effect of feature selection on the accuracy of emulation. Based on our previous work [35] and following the description in II-D, we created four GP emulators with configurations summarized in Tab. III. The dimensionality reduction was fixed to 10 PCA components. We expect that a feature selection (configurations #2 and #4) will achieve accuracy between the Gaussian kernel and ARD-Gaussian kernel emulators without feature selection (configurations #1 and #3 respectively).

TABLE III
MAIN GP EMULATORS CONFIGURATION PARAMETERS.

ID	Kernel	Feat. selection (χ_k)
#1	Gaussian	No (–)
#2	Gaussian	Yes (L_1)
#3	ARD-Gaussian	No (–)
#4	Mahalanobis Gaussian	Yes (L_1)

A vegetation Lambertian surface reflectance, $\boldsymbol{\rho}$, was propagated by Eq. (16) to top-of-atmosphere (TOA) using the transfer functions in the *reference* dataset. This resulted in m TOA radiance spectra, $\{\mathbf{L}\}_{i=1}^{m}$:

$$\mathbf{L}_i = \mathbf{L}_{0,i} + \frac{(\mathbf{E}_{dir,i}\mu_{il,i} + \mathbf{E}_{dif,i})(\mathbf{T}_{dir,i} + \mathbf{T}_{dif,i})\boldsymbol{\rho}}{\pi(1-\mathbf{S}_i\boldsymbol{\rho})}, \tag{16}$$

where $\mu_{il,i} = \cos(SZA_i)$. Here, the sub-index i identifies a sample with input conditions $\mathbf{x}_i$. Each emulator was run to predict the transfer functions given the input conditions of the *reference* dataset, after which they were used to invert the surface reflectance from (16). As a result, a dataset of m surface reflectance spectra $\{\widehat{\boldsymbol{\rho}}\}_{i=1}^{m}$ was generated. We calculated the relative difference against the reference surface reflectance $\boldsymbol{\rho}$ at every wavelength, i.e., $\boldsymbol{\varepsilon}_i = (\widehat{\boldsymbol{\rho}}_i - \boldsymbol{\rho}_i)/\boldsymbol{\rho}_i$, and the mean relative error (MRE) over the entire *reference* dataset, i.e., MRE $= \frac{1}{m}\sum_{i=1}^{m} \varepsilon_i$ to better visualize and compare the accuracy of all emulators. We also calculated the spectrally-averaged MRE (MRE$_\lambda$), excluding wavelengths in the saturating H2O bands (1325-1500 nm and 1750-1950 nm). The relative error histogram at every wavelength and the emulator runtime was also kept for further analysis.

All the analyses were carried out on a personal computer with the following specifications: Windows 10 64-bit OS, i7-4710 CPU 2.50 GHz, 16 GB RAM, and 12 CPUs.

IV. Results

A. Feature ranking and GSA results

We start by presenting the results that analyze the influence of the cost function on the performance of the feature ranking method. The $V(k)$ curve was calculated at four selected wavelengths and with four cost functions. The values of $V(k)$ are normalized for their corresponding values at k=1 to better compare the results from each cost function. The results are presented only for two transfer functions (E_{dif} and T_{dir}) to illustrate the main effects of the various features in the radiative transfer. For the diffuse solar irradiance (see Fig. 3) we observe a similar behavior of the normalized $V(k)$ curves for all the tested cost function. In all cases, $V(k)$ shows a decreasing trend with the biggest decrease happening when the number of features is increased from k=1 to k=2. All curves show a nearly flat behavior after 4 to 6 features, indicating that additional features only contain residual information. The results for the direct transmittance are shown in Fig. 4. Again, all cost functions obtain a similar $V(k)$ curve and a flat behavior after 3 to 4 features, depending on the wavelength.

Based on these curves, the SIC method automatically selects the number of relevant features. Tab. IV displays the number of relevant features for all cost functions and wavelengths. For the diffuse solar irradiance, the selected features range from 4 to 6, depending on the wavelength and chosen cost function. For each wavelength, the number of selected features remains consistent at 400 nm and at 761 nm. At 603 nm, the number oscillates between 4 (with L_1, L_1^r, and L_2^r) and 6 (with L_2). At 940 nm, the selected features range from 5 to 6. For the direct transmittance, the results indicate that the forward selection method is insensitive to the cost function, with a number of selected features between 3 (at 400 nm) and 5 (at 603 nm).

To complement these results, we conducted a regression-based GSA using the $V(k)$ curves to calculate Sensitivity Indices (SI). Figs. 5 and 6 depict the GSA results for E_{dif} and T_{dir} across

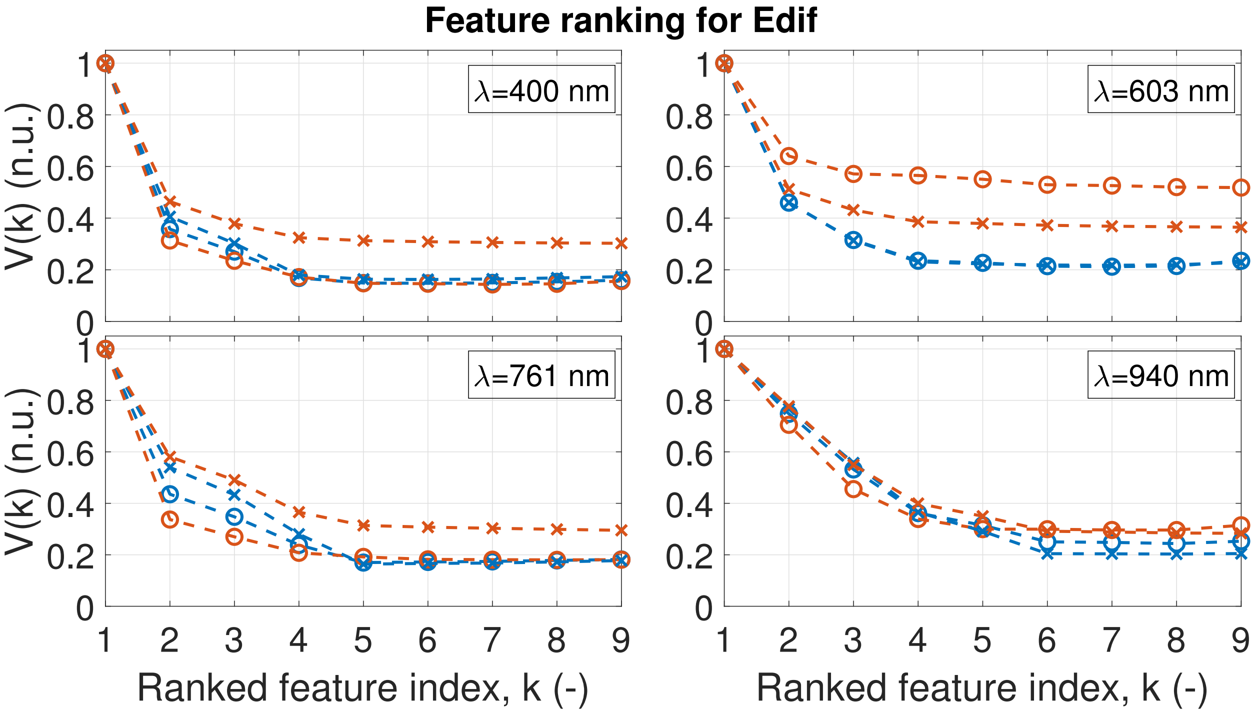

Fig. 3. Curve $V(k)$, normalized to 1 at k=1, at various wavelengths (λ) and cost functions: L_1 (blue) and L_2 (red). Markers indicate absolute, ○, and relative, x, norms. Results for at-surface diffuse solar irradiance (E_{dif}).

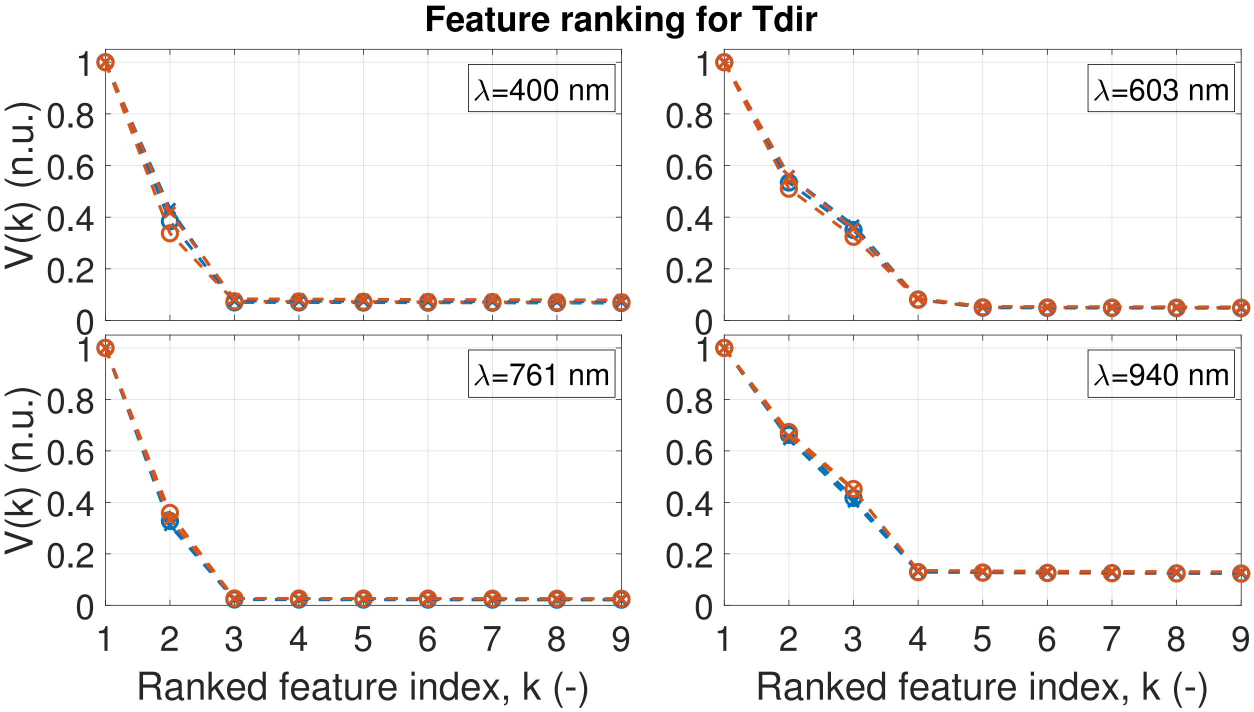

Fig. 4. Same as Fig. 3 but for direct transmittance (T_{dir}).

TABLE IV

Number of selected features for E_{dif} and T_{dir} at a confidence level ℓ=98%. L_p^r refers to the relative L_p norm.

	E_{dif}				T_{dir}			
	L_1	L_2	L_1^r	L_2^r	L_1	L_2	L_1^r	L_2^r
400 nm	5	5	5	5	3	3	3	3
603 nm	4	6	4	4	5	5	5	5
761 nm	5	5	5	5	3	3	3	3
940 nm	6	5	6	6	4	4	4	4

various wavelengths and cost functions. Each color in the figures corresponds to an input feature (see Tab. II), and the bar size reflects its importance (SI). For E_{dif} (see Fig. 5), we observe that all cost functions get similar GSA results. Two main features (SZA 🟥 and AOT 🟦) are always identified as the key drivers for all wavelengths with a SI of 30% to 80% for AOT and 15% to 60% for SZA. The relative norms show more sensitivity to secondary features (surface elevation 🟧, α 🟩, and g 🟩) with an SI ranging from 5% to 20% depending on the wavelength. The results from the various wavelengths indicate that the forward selection method is sensitive to the expected relevant features. For example, at 761 nm the method shows a higher relevance of the surface elevation, and at 940 nm it is sensitive to CWV 🟦. The RAA 🟥 does not show any influence in the GSA results, in line with the expected for nadir-view simulations. For T_{dir} (see Fig. 6), we observe very similar GSA results regardless of the cost function. Here, the leading features are AOT, α, and surface elevation. At 603 nm we observe a small influence of the O3 ⬛. The SZA, RAA, and scattering-related features (g) do not have any influence.

The regression-based GSA results are compared with two independent methods: the variance-based GSA method proposed in [12], and a method based on the scale-length hyper-parameters of an ARD-Gaussian GP regressor (see Fig. 7). The variance-based GSA method yields SI values for E_{dif} that closely align with those obtained by the regression-based GSA, particularly with the L_1 and L_2 absolute norms. However, for T_{dir}, the values slightly differ from the regression-based GSA. Specifically, at 400 nm, the variance-based assigns more importance to AOT (80%) compared to the regression-based method (65-70%). At 761 nm, surface elevation is the dominating feature with an SI of 60% according to the variance method, while the regression method only attributes 40%. Similarly, at 940 nm, CWV also holds higher importance (50%) than in the regression-based results (30%). Nonetheless, the key features and ranking are consistent

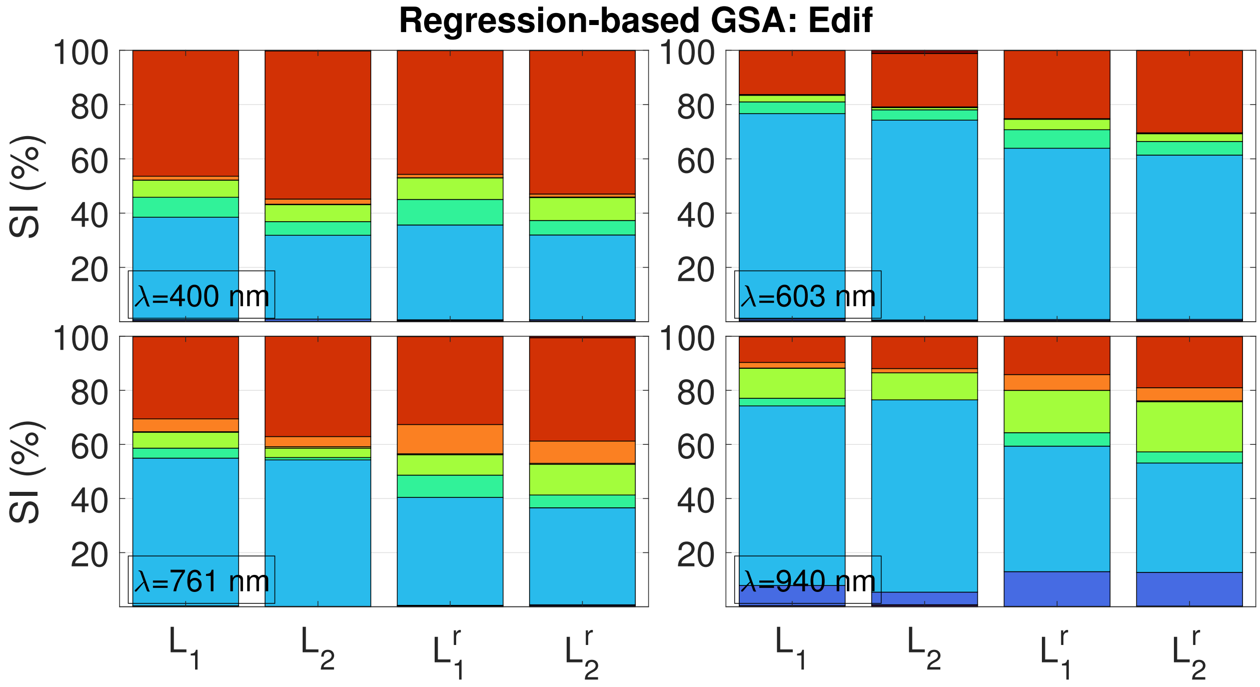

Fig. 5. GSA results for E_{dif} and each tested cost function. L_p^r refers to the relative L_p norm. See Tab. II to identify each feature from its color.

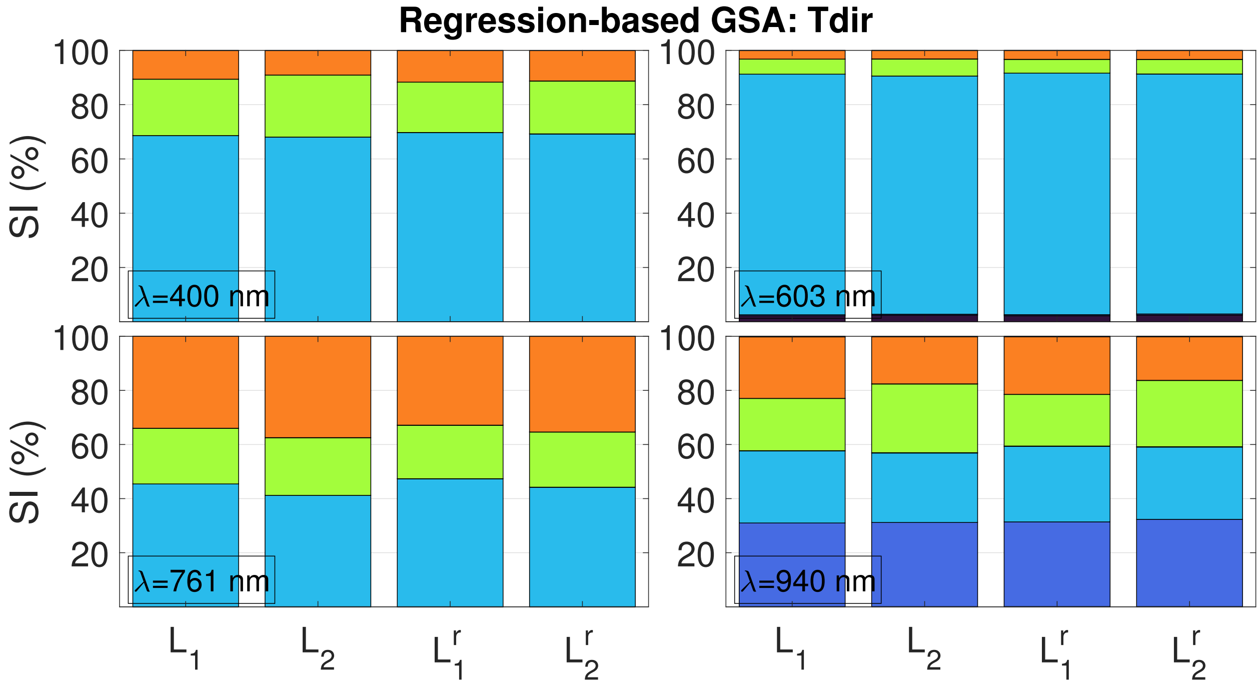

Fig. 6. Same as Fig. 5 but for T_{dir}.

between the two methods. The ARD-based results consistently yield SI values for the T_{dir}, except at 940 nm where CWV gains higher importance (50%). For the E_{dif}, the results are also consistent with the regression-based GSA but it shows higher relevance for features of secondary importance, particularly at 940 nm.

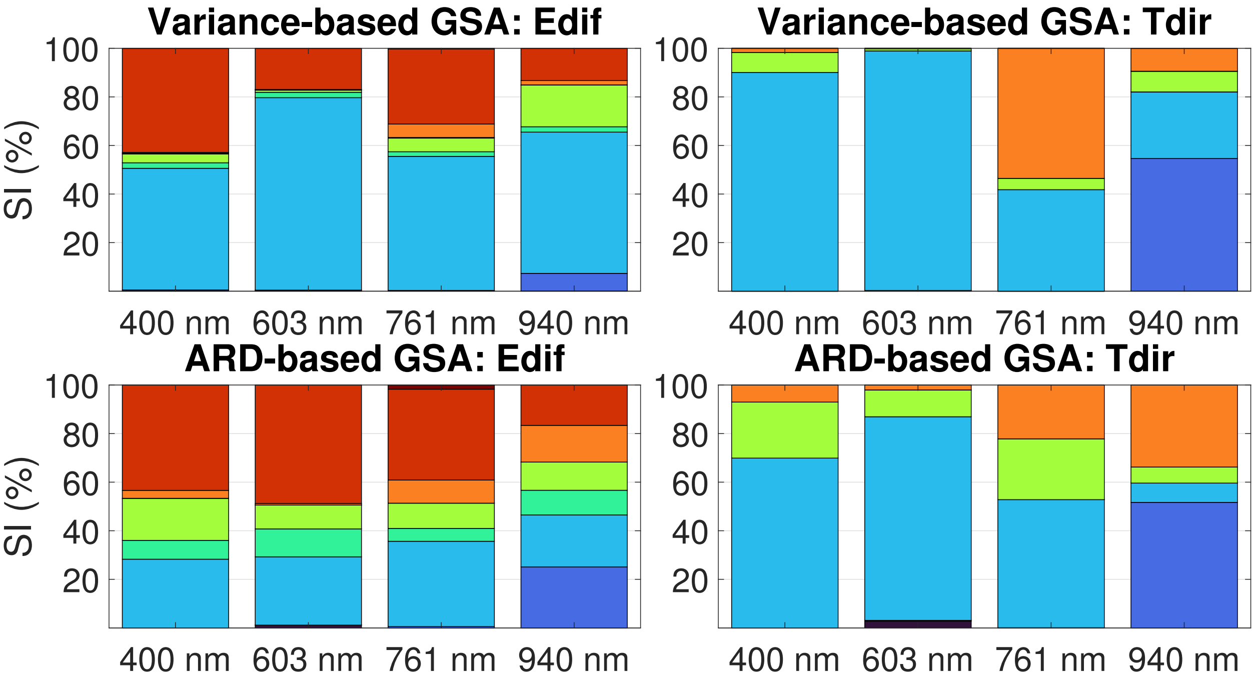

Fig. 7. GSA results for E_{dif} (left) and T_{dir} obtained from a variance-based method (top) and the ARD-Gaussian kernel hyperparameters (bottom). See Tab. II to identify each feature from its color.

Next, we show the GSA results for the entire spectral range (400-2500 nm) at 5 cm^{-1} resolution. Fig. 8 compares the SI obtained with our regression-based method (left column) against the results from the variance-based method (right column). The results are displayed for three transfer functions (path radiance, at-surface diffuse irradiance, and direct transmittance) to illustrate the main effects of the input features on the radiative transfer. In general, both methods produce nearly indistinguishable results. In all displayed transfer functions, the driving features outside of H_2O absorptions are the AOT, SZA, α, and g. Surface elevation is also a key feature within absorption regions such as H_2O and, particularly, inside the O_2-A band (760-770 nm). In terms of SI, the O_3 absorption has a residual effect (SI<3% given the large variability of atmospheric conditions in the simulated dataset. The biggest differences in GSA results occur within the deep H_2O bands at 1325-1500 nm and 1750-1950 nm due to near-zero radiances and transmittances resulting from water vapor absorption saturation. Numerical errors are thus prominent in these bands and derive nonphysical results such as a sensitivity to O_3.

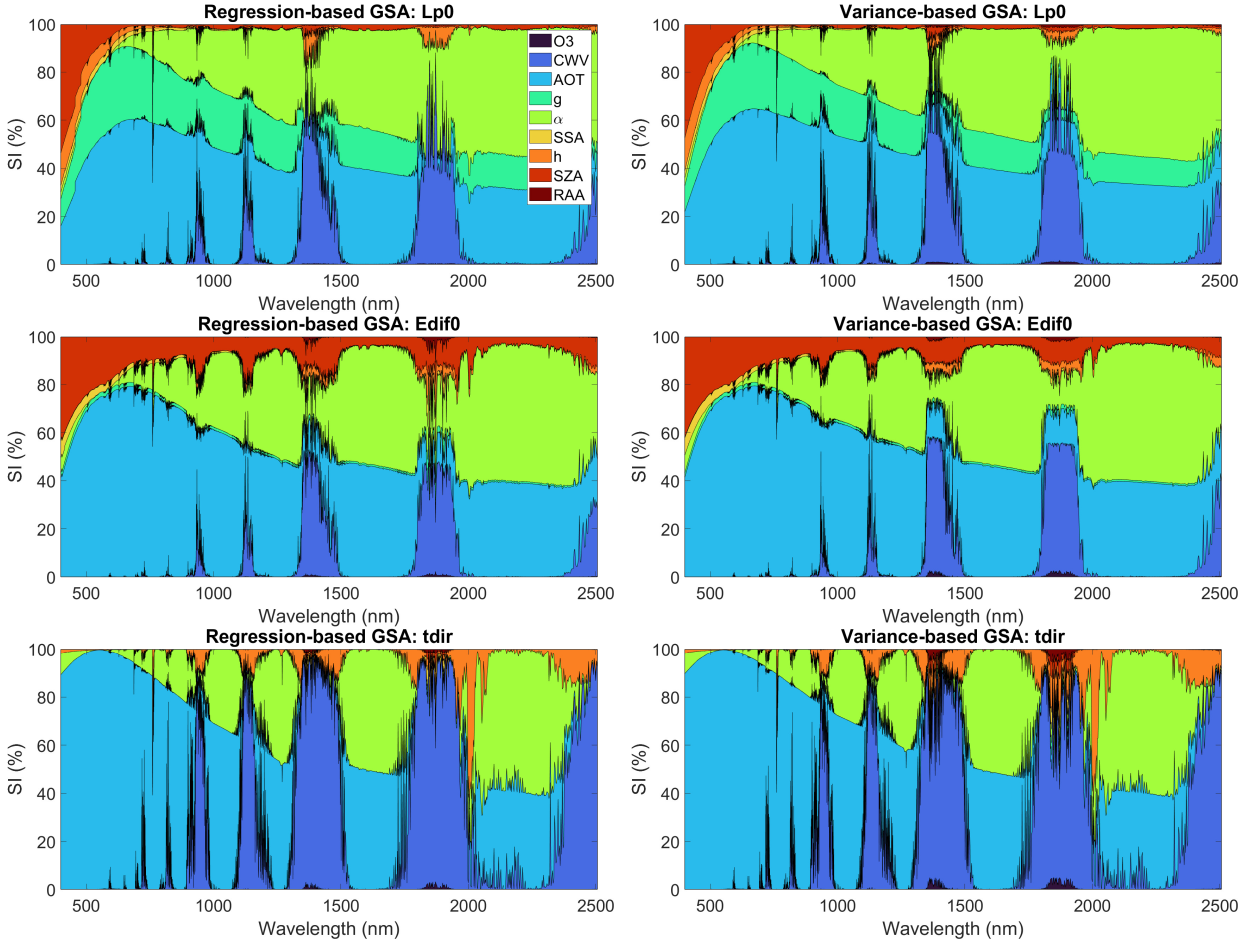

Fig. 8. GSA results from the proposed regression-based feature ranking method (left) and variance-based method [12] (right). Path radiance (top), at-surface diffuse solar irradiance (middle), and upwards direct transmittance (bottom).

B. Emulation results

We analyzed how feature selection impacts the accuracy of an emulator. For that, we compared the performance of the GP emulators in Tab. III by plotting the spectral mean relative errors (MRE) (see Fig. 9). All emulators show similar spectral behavior of the MRE, where the higher values correspond to spectral regions with lower surface reflectance values, as expected given the nature of the relative error. In addition, the errors are higher inside gaseous absorption (mainly H_2O and O_2) due to divisions by nearly zero during the inversion of surface reflectance. Moreover, the MRE values tend to be higher towards shorter wavelengths (<500 nm) due to the impact of aerosol scattering (see Fig. 8). In terms of accuracy, the highest errors are obtained with the basic GP emulator (Gaussian kernel and no feature selection, configuration #1). Feature selection (configuration #2) improves the results by 0.1% to 0.2% depending on the wavelength.

The only exception is inside the O_3 absorption (530-640 nm), where the GP emulator with feature selection obtains the highest errors (1.5%). Applying the feature ranking through the Mahalanobis kernel (configuration #4), the results are improved by nearly a factor 2 in the visible spectral range (400-700 nm) and remain the same in the rest of the spectral range. The emulator with an ARD-Gaussian kernel (configuration #3) achieves the lowest errors in all wavelengths with MRE values 0.2% to 0.6% outside of absorption bands.

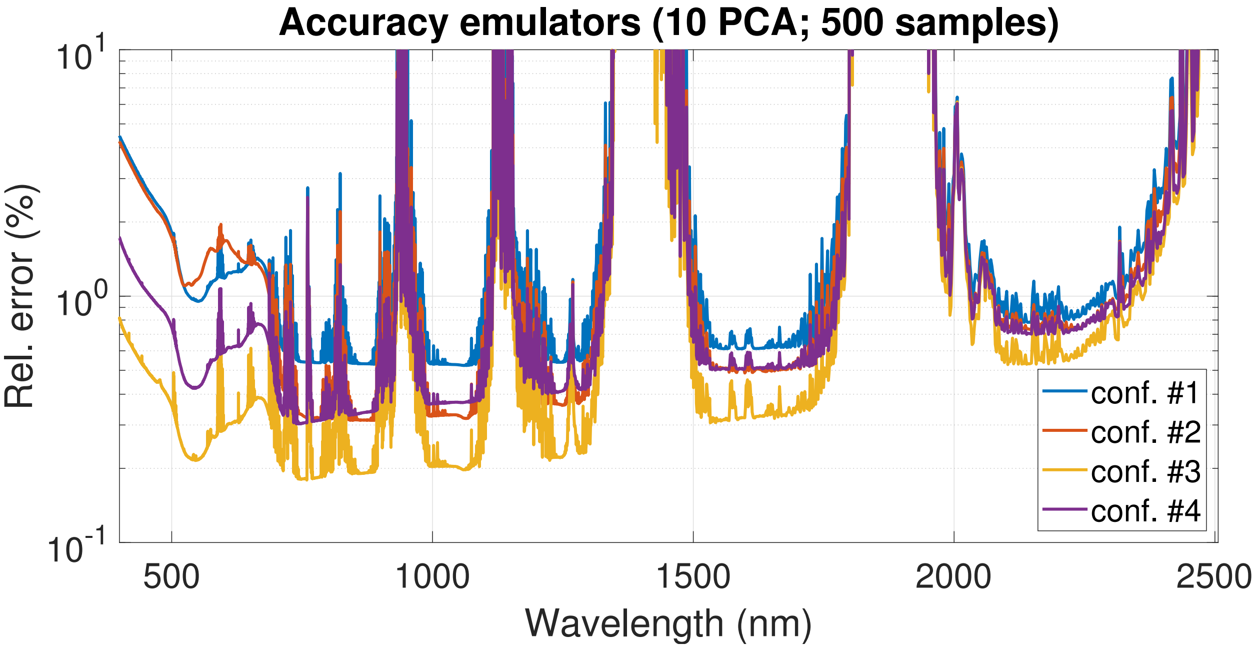

Fig. 9. Spectral MRE (in %) for various emulators with and without feature selection (see legend).

Tab. V summarizes the accuracy results from the three GP emulator configurations, along with their runtimes when predicting 10000 samples. These results demonstrate that feature selection enhances the accuracy of a GP emulator with a Gaussian kernel by 0.2%, all while maintaining a negligible increase in runtime. Nonetheless, the ARD-Gaussian GP emulator exhibits superior accuracy, albeit at the cost of longer runtime. The Mahalanobis Gaussian kernel (configuration #4) requires as much runtime as the ARD-Gaussian GP emulator but with nearly twice the value of MRE_λ. The training time of all these emulators (including feature selection) is within a few minutes, the slowest being for the Mahalanobis Gaussian kernel emulator.

TABLE V
ACCURACY (MRE$_\lambda$, IN %), RUNTIME (IN S), AND EXPLAINABILITY FOR THE PREDICTION OF 10000 SAMPLES USING THE GP EMULATOR CONFIGURATION IN TAB. III.

Config.:	**#1**	**#2**	**#3**	**#4**
MRE$_\lambda$ (%)	1.8	1.6	0.5	0.9
Runtime (s)	1.4	1.5	2.3	2.3
Explainability	None	Medium	Low	High

V. DISCUSSION

A. *On the feature selection method*

Model explainability is of paramount importance in Earth Observation as it enables scientists and decision-makers to understand how input features impact model outputs, gaining insights into the processes and variables driving observations from remote sensing satellites. While deterministic RTMs provide explainable results through the laws of physics implemented in the underlying mathematical code, statistical regression models are often perceived as black boxes that derive trends from the analysis of data. Explainability ensures that statistical models remain accessible, transparent, and accountable, ultimately contributing to more accurate remote sensing data processing algorithms. In this context, we introduce a method for automated feature ranking and selection, which not only enhances model explainability through GSA but also holds the potential to improve regression models, such as emulation. While our method was initially applied to atmospheric RTM simulations, it also applies to a broader range of multi-dimensional models.

In our application, both wrapper and filter feature selection methods are deemed suitable for global sensitivity analysis and emulation. We opted for a wrapper-forward selection algorithm [10] striking a balance between accuracy and runtime. This algorithm recursively adds the most relevant features to a regression model to minimize a cost function. By doing so, it accounts for the combined impact of features on the model's performance and it is well-suited for optimizing emulator accuracy. In contrast, filter methods, although faster, rely on ad hoc statistical criteria that may not be ideal for our regression optimization task. On the other hand, intrinsic feature selection methods are based on machine learning algorithms lacking transparency in explicitly revealing the relationship between input features and RTM outputs. This lack of direct interpretability might limit insights into the individual impact of each feature on the model's outcome. Alternatively, a wrapper-backward selection algorithm could have been used to add less relevant features by maximizing a cost function. Another option is to remove one feature at a time. Nonetheless, we anticipate that the feature ranking will remain relatively consistent when using these alternative wrapper algorithms because some input features have little to no impact on the model's performance (e.g., RAA for nadir observations). Adding or removing such features should not significantly alter the performance of the method, and all methods should yield similar feature rankings. This is partly because the cost functions produce similar values regardless of the method used, which may result in similar selections. Additionally, when the cost function is symmetric in terms of adding or removing features, it can lead to convergence in results. However, it is worth noting that these alternative feature ranking methods might produce different outcomes for less stable performance regression models, where adding or removing a single feature could significantly impact the model's accuracy.

The behavior of the feature selection method depends on the selected regression method and cost function. For the underlying regression method, we opted for polynomial fitting as it captures with sufficient accuracy the main relationships between the RTM input features and output spectral data. While we tested GP regression as an alternative (results not shown), we found no substantial change in the feature ranking results. Yet, for other models with more complex input-output relationships, advanced regression methods may be necessary. As for the cost function, it evaluates the importance of each input feature in the model output. Although the choice of the cost function can influence the results, we observed that for single-output scenarios, its impact is residual in terms of the number of selected features and the relative influence of each feature. Nevertheless, for multi-output (spectral) data, we expect that the cost function will have a bigger influence. For instance, relative norms can potentially over-influence the information from spectral channels within deep H_2O absorptions. Although this can be seen as a drawback, it could be converted to a benefit since various rankings can be obtained using different cost functions. This would allow tailoring the cost function for extracting specific information from the data. An illustrative example is when we are particularly interested in processing low radiance values, as is the case with O_2 absorption (e.g., FLEX mission [45]). Here, the use of relative norms as the cost function might prove to be beneficial. The choice of the optimal number of the most relevant features is based on the SIC method. SIC generalizes and unifies several information criterion approaches given in the literature such as [26], [27], containing them as special cases in the set of Eq. (8). Moreover, it has been shown in [24] that the choice in Eq. (9) is more robust than using other information criteria.

B. On the Global Sensitivity Analysis

We initially assessed the suitability of the presented featured selection method in the context of GSA. GSA offers a comprehensive perspective for discerning the relative influence of each input feature on model output [28]. Our regression-based GSA method yields physically meaningful results across all wavelengths. Specifically, (1) the scattering variables exclusively affect the diffuse components of radiation, (2) the SZA solely impacts downward irradiance while RAA has no influence due to the simulations being carried out at nadir, and (3) gas concentrations (O_3 and H_2O) and surface elevation exert a more pronounced influence within absorption spectral regions. Our regression-based GSA results consistently align with those obtained from a fundamentally different variance-based GSA method [44], [46], thus validating the robustness of our approach. However, it is noted that the GSA results yield unrealistic results in saturating H_2O absorbtion bands (e.g., influence of O_3 concentration, lower relevance of H_2O concentration) due to the enhancement of numerical errors. Furthermore, we compared our GSA findings with the information provided by the hyperparameters of an ARD-Gaussian GP regressor. While the ARD-based GSA outcomes identified the same key features as our method, there were discrepancies in the magnitude and ranking of the sensitivity indices. Notably, the ARD-based approach assigned higher sensitivity to less important variables in comparison to the regression- and variance-based GSA methodologies. This might be explained because the ARD-Gaussian GP model focuses on optimizing hyperparameters to fit the observed data and enhance the predictive accuracy of the regression model.

C. On the application for emulation

We then incorporated our feature selection method into GP emulators to enhance the models' interpretability by identifying the key features influencing the model outputs. We expected that this would lead to improved predictive accuracy for GP emulators. We developed two distinct approaches to integrate feature selection within the emulator framework. The first approach (configuration #2) involves selecting the key features directly from the multi-output (spectral) data. This approach is straightforward and efficient, resulting in a modest reduction of 0.2% in the relative errors compared to the standard GP emulator (configuration #1) while maintaining a similar runtime. However, it is worth noticing that the accuracy improvement varies across the spectral range, with higher errors observed in the O_3 region. This is due to the multi-output feature selection method excluding the O_3 feature when considering all wavelengths. Choosing a larger confidence interval in the SIC method should yield better results. The second approach (configuration #4, called Mahalabobis Gaussian kernel) involves applying sensitivity indices to re-scale each input dimension. Conceptually, this approach is akin to an ARD-Gaussian GP emulator (configuration #3) and achieves the same runtime. Although this approach delivers a notable improvement in predictive accuracy (MRE_λ=0.9%), especially at shorter wavelengths, it still does not reach the accuracy of the ARD-Gaussian emulator (MRE_λ=0.5%). This lower accuracy is associated with observed differences when comparing the regression- and ARD-based GSA methods. Unlike the ARD-Gaussian kernel, the Mahalabobis Gaussian kernel prioritizes the physical explainability of the emulator hyperparameters at the expense of predictive accuracy. Nonetheless, we suggest that the Mahalanobis Gaussian method could serve to initialize the hyperparameters optimization in an ARD-Gaussian kernel. Furthermore, we anticipate that this method may potentially achieve superior accuracy in the regression of biophysical variables. Indeed, this method would be particularly beneficial when dealing with hyperspectral surface reflectance data, where optimizing an ARD-Gaussian GP regression may become unfeasible due to the large number of hyperparameters [47].

VI. Conclusions

We developed an automated feature selection method to construct physics-aware emulators and improve their explainability. Our method involved the combination of a wrapper forward selection algorithm, generalized for multioutput models, with the *spectral information criterion*. This methodology was applied to spectral data generated by an atmospheric RTM and validated against two independent methods: variance-based GSA and ARD. Our GSA results indicate consistent sensitivity indices with the variance-based GSA across all wavelengths. However, the ARD results differ from the other two GSA approaches, attributing higher sensitivity to secondary features. We interpret these results as suggesting that the ARD version prioritizes the reduction of prediction errors in a GP regressor over emphasizing physical explainability. In contrast, the regression- and variance-based GSA approaches offer a more comprehensive physical explanation of the RTM data. We then applied our feature selection method to create two different implementations of a physics-aware emulator. The first involves direct feature selection on the spectral data (configuration #2), while the second entails constructing a Mahalanobis Gaussian kernel on the GP regressor for each PCA component (configuration #4). We compared the accuracy of these two physics-aware emulators against a standard emulator based on GP regression with a Gaussian kernel and its ARD extension (configurations #1 and #3). Our results reveal that physics-aware emulation through feature selection improves accuracy by 0.2%, achieved by eliminating features with low relevance in the model and enhancing model explainability. Nonetheless, the ARD-Gaussian kernel GP emulator achieves the highest prediction accuracy in the entire spectral range. While the feature selection presented here yields only a modest improvement in emulation accuracy, we discuss its potential for regression (inversion) of biophysical parameters from hyperspectral surface reflectance data. We consider that the proposed methodology can overcome the limitations of an ARD-Gaussian GP regression in very high-dimensional input spaces, eliminating the need for dimensionality reduction and improving physical-awareness of the retrieval algorithm. We expect that the feature selection method presented here will contribute to the development of advanced physics-aware emulators, leading to improved accuracy and performance while facilitating the interpretability of results when processing satellite data.

References

[1] G. Camps-Valls, J. Verrelst, J. Muñoz-Marí, V. Laparra, F. Mateo-Jiménez, and J. Gómez-Dans, "A survey on Gaussian processes for Earth Observation Data Analysis," *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 2, 2016.

[2] J. Verrelst, L. Alonso, G. Camps-Valls, J. Delegido, and J. Moreno, "Retrieval of vegetation biophysical parameters using Gaussian process techniques," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 5 PART 2, pp. 1832–1843, 2012.

[3] G. Camps-Valls, D. Svendsen, L. Martino, J. Muñoz Marí, V. Laparra, M. Campos-Taberner, and D. Luengo, "Physics-aware Gaussian processes in remote sensing," *Applied Soft Computing*, vol. 68, pp. 69–82, Jul 2018.

[4] J. Cortés-Andrés, G. Camps-Valls, S. Sippel, E. Székely, D. Sejdinovic, E. Diaz, A. Pérez-Suay, Z. Li, M. Mahecha, and M. Reichstein, "Physics-aware Nonparametric Regression Models for Earth Data Analysis," *Environmental Research Letters*, vol. 17, no. 5, 2022.

[5] L. Li, J.-F. Wang, M. Franklin, Q. Yin, J. Wu, G. Camps-Valls, Z. Zhu, C. Wang, Y. Ge, and M. Reichstein, "Improving air quality assessment using physics-inspired deep graph learning," *npj Climate and Atmospheric Science*, 2023.

[6] M. Reichstein, G. Camps-Valls, B. Stevens, J. Denzler, N. Carvalhais, M. Jung, and Prabhat, "Deep learning and process understanding for data-driven Earth System Science," *Nature*, vol. 566, pp. 195–204, Feb. 2019.

[7] L. Gomez-Chova, G. Camps-Valls, J. Calpe-Maravilla, L. Guanter, and J. Moreno, "Cloud-screening algorithm for envisat/meris multispectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 12, pp. 4105–4118, 2007.

[8] S. Mishra and R. Molinaro, "Physics informed neural networks for simulating radiative transfer," *Journal of Quantitative Spectroscopy and Radiative Transfer*, vol. 270, p. 107705, 2021.

[9] J. Verrelst, Z. Malenovsky, C. Van der Tol, G. Camps-Valls, J.-P. Gastellu-Etchegorry, P. Lewis, P. North, and J. Moreno, "Quantifying vegetation biophysical variables from imaging spectroscopy data: A review on retrieval methods," *Surveys in Geophysics*, vol. 40, no. 3, pp. 589–629, 2019.

[10] G. Camps-Valls, J. Mooij, and B. Schölkopf, "Remote sensing feature selection by kernel dependence measures," *IEEE Geoscience and Remote Sensing Letters*, vol. 7, no. 3, pp. 587–591, 2010.

[11] L. Martino and J. Read, “A joint introduction to Gaussian Processes and Relevance Vector Machines with connections to Kalman filtering and other kernel smoothers,” *Information Fusion*, vol. 74, pp. 17–38, 2021.

[12] A. Saltelli, P. Annoni, I. Azzini, F. Campolongo, M. Ratto, and S. Tarantola, “Variance based sensitivity analysis of model output. design and estimator for the total sensitivity index,” *Computer Physics Communications*, vol. 181, no. 2, pp. 259–270, 2010.

[13] R. M. Neal, *Bayesian Learning for Neural Networks*, 1st ed., ser. Lecture Notes in Statistics. New York, NY: Springer, 1996, vol. 118.

[14] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. New York: The MIT Press, 2006.

[15] G. Camps-Valls, D. Tuia, X. Zhu, and M. Reichstein, *Deep learning for the Earth Sciences: A comprehensive approach to remote sensing, climate science and geosciences*. Wiley & Sons, 2021.

[16] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[17] G. Tramontana, K. Ichii, G. Camps-Valls, E. Tomelleri, and D. Papale, “Uncertainty analysis of gross primary production upscaling using Random Forests, remote sensing and eddy covariance data,” *Remote Sensing of Environment*, vol. 168, pp. 360–373, 2015.

[18] J. Verrelst, J. Rivera, A. Gitelson, J. Delegido, J. Moreno, and G. Camps-Valls, “Spectral Band Selection for Vegetation Properties Retrieval using Gaussian Processes Regression,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 52, pp. 554–567, 2016.

[19] Z. Chen, D. Hong, and H. Gao, “Grid Network: Feature Extraction in Anisotropic Perspective for Hyperspectral Image Classification,” *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.

[20] A. O’Hagan, “Bayesian analysis of computer code outputs: A tutorial,” *Reliability Engineering and System Safety*, vol. 91, no. 10-11, pp. 1290–1300, 2006.

[21] J. L. Gómez-Dans, P. E. Lewis, and M. Disney, “Efficient emulation of radiative transfer codes using Gaussian processes and application to land surface parameter inferences,” *Remote Sensing*, vol. 8, no. 2, p. 119, 2016.

[22] J. Verrelst, J. P. Rivera Caicedo, J. Muñoz-Marí, G. Camps-Valls, and J. Moreno, “Scope-based emulators for fast generation of synthetic canopy reflectance and sun-induced fluorescence spectra,” *Remote Sensing*, vol. 9, no. 9, p. 927, 2017.

[23] J. Vicent, J. Verrelst, J. Rivera Caicedo, N. Sabater Medina, J. Muñoz, G. Camps-Valls, and J. Moreno, “Emulation as an accurate alternative to interpolation in sampling radiative transfer codes,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 12, pp. 1–14, 10 2018.

[24] L.Martino, R. S. Millan-Castillo, and E. Morgado, “Spectral information criterion for automatic elbow detection,” *Expert Systems with Applications*, vol. 231, p. 120705, 2023.

[25] P. Stoica and Y. Selén, “Model-order selection: a review of information criterion rules,” *IEEE Signal Processing Magazine*, pp. 36–47, 2004.

[26] G. Schwarz *et al.*, “Estimating the dimension of a model,” *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[27] D. Spiegelhalter, N. G. Best, B. P. Carlin, and A. V. der Linde, “Bayesian measures of model complexity and fit,” *J. R. Stat. Soc. B*, vol. 64, pp. 583–616, 2002.

[28] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola, *Global sensitivity analysis: the primer*. UK: John Wiley & Sons, Dec. 2008.

[29] J. Verrelst, J. Rivera, C. Van Der Tol, M. F., G. Mohammed, and J. Moreno, “Global sensitivity analysis of the scope model: what drives simulated canopy-leaving sun-induced fluorescence?” *Remote Sensing of Environment*, vol. 166, pp. 8–21, 2015.

[30] G. Heinze, C. Wallisch, and D. Dunkler, “Variable selection - a review and recommendations for the practicing statistician,” *Biometrical journal*, vol. 60, no. 3, pp. 431–449, 2018.

[31] R. S. Millan-Castillo, L. Martino, E. Morgado, and F. Llorente, “An exhaustive variable selection study for linear models of soundscape emotions: Rankings and Gibbs analysis,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2460–2474, 2022.

[32] A. Berk, G. Anderson, P. Acharya, L. Bernstein, L. Muratov, J. Lee, M. Fox, S. Adler-Golden, J. Chetwynd, M. Hoke, R. Lockwood, J. Gardner, T. Cooley, C. Borel, P. Lewis, and E. Shettle, “MODTRAN$^{\text{TM}}$5: 2006 update,” *Proceedings of SPIE - The International Society for Optical Engineering*, vol. 6233 II, 2006.

[33] P. G. Brodrick, D. R. Thompson, J. E. Fahlen, M. L. Eastwood, C. M. Sarture, S. R. Lundeen, W. Olson-Duvall, N. Carmon, and R. O. Green, “Generalized radiative transfer emulation for imaging spectroscopy reflectance retrievals,” *Remote Sensing of Environment*, vol. 261, p. 112476, 2021.

[34] J. Vicent Servera, J. P. Rivera-Caicedo, J. Verrelst, J. Muñoz-Marí, N. Sabater, B. Berthelot, G. Camps-Valls, and J. Moreno, “Systematic Assessment of MODTRAN Emulators for Atmospheric Correction,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2022.

[35] J. Vicent Servera, L. Martino, J. Verrelst, and G. Camps-Valls, “Multifidelity Gaussian Process Emulation for Atmospheric Radiative Transfer Models,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–10, 2023.

[36] R. L. Thorndike, “Who belongs in the family?” *Psychometrika*, vol. 3, pp. 267–276, 1953.

[37] P. Stoica and Y. Selén, “Cross-validation rules for order estimation,” *Digital Signal Processing*, vol. 14, pp. 355–371, 2004.

[38] M. Efroymson, “Multiple regression analysis,” *Mathematical methods for digital computers*, pp. 191–203, 1960.

[39] R. R. Hocking, “The analysis and selection of variables in linear regression,” *Biometrics*, pp. 1–49, 1976.

[40] F. Llorente, L. Martino, D. Delgado, and J. Lopez-Santiago, “Marginal likelihood computation for model selection and hypothesis testing: an extensive review,” *SIAM Review*, vol. 65, no. 1, pp. 3–58, 2023.

[41] E. J. Hannan and B. G. Quinn, “The determination of the order of an autoregression,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 41, no. 2, pp. 190–195, 1979.

[42] E. Morgado, L. Martino, and R. S. Millan-Castillo, “Universal and automatic elbow detection for learning the effective number of components in model selection problems,” *Digital Signal Processing*, vol. 140, p. 104103, 2023.

[43] M. McKay, R. Beckman, and W. Conover, “Comparison of three methods for selecting values of input variables in the analysis of output from a computer code,” *Technometrics*, vol. 21, no. 2, pp. 239–245, 1979.

[44] J. Vicent, J. Verrelst, N. Sabater, L. Alonso, J. P. Rivera-Caicedo, L. Martino, J. Muñoz-Marí, and J. Moreno, “Comparative analysis of atmospheric radiative transfer models using the Atmospheric Look-up table Generator (ALG) toolbox (version 2.0).” *Geoscientific Model Development*, vol. 13, no. 4, 2020.

[45] M. Drusch and et al., “The FLuorescence EXplorer Mission Concept-ESA’s Earth Explorer 8,” *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1 – 12, 2016.

[46] J. Verrelst, N. Sabater, J. P. Rivera, J. Muñoz Marí, J. Vicent, G. Camps-Valls, and J. Moreno, “Emulation of leaf, canopy and atmosphere radiative transfer models for fast global sensitivity analysis,” *Remote Sensing*, vol. 8, no. 8, p. 673, 2016.

[47] J. Gómez, J. Blasco, E. Molto, and G. Camps-Valls, “Hyperspectral detection of citrus damage with mahalanobis kernel classifier,” *Electronics Letters*, vol. 43, pp. 1082–1084, 2007.