ML-AMPSIT: Machine Learning-based Automated Multi-method Parameter Sensitivity and Importance analysis Tool

Dario Di Santo[1], Cenlin He[2], Fei Chen[3], and Lorenzo Giovannini[1]

[1]Department of Civil, Environmental and Mechanical Engineering, University of Trento, Trento, Italy
[2]NSF National Center for Atmospheric Research (NCAR), Boulder, CO, USA
[3]Division of Environment and Sustainability, Hong Kong University of Science and Technology, Hong Kong, China

Correspondence: Dario Di Santo (dario.disanto@unitn.it)

Abstract. The accurate calibration of parameters in atmospheric and Earth system models is crucial for improving their performance, but remains a challenge due to their inherent complexity, which is reflected in input-output relationships often characterized by multiple interactions between the parameters and thus hindering the use of simple sensitivity analysis methods. This paper introduces the Machine Learning-based Automated Multi-method Parameter Sensitivity and Importance analysis Tool (ML-AMPSIT), a new tool designed with the aim of providing a simple and flexible framework to estimate the sensitivity and importance of parameters in complex numerical weather prediction models. This tool leverages the strengths of multiple regression-based and probabilistic machine learning methods including LASSO, Support Vector Machine, Classification and Decision Trees, Random Forest, Extreme Gradient Boosting, Gaussian Process Regression, and Bayesian Ridge Regression. These regression algorithms are used to construct computationally inexpensive surrogate models to effectively predict model outputs from input parameters, thereby significantly reducing the computational burden of running high-fidelity models for sensitivity analysis. Moreover, the multi-method approach allows for a comparative analysis of the results. Through a detailed case study with the Weather Research and Forecasting (WRF) model coupled with the Noah-MP land surface model, ML-AMPSIT is demonstrated to efficiently predict the behavior of Noah-MP model parameters with a relatively small number of model runs, by simulating a sea breeze circulation over an idealized flat domain. This paper points out how ML-AMPSIT can be an efficient tool for performing sensitivity and importance analysis also for complex models, guiding the user through the different steps and allowing for a simplification and automatization of the process.

1 Introduction

One of the primary sources of error in atmospheric/Earth system models stems from inaccurate parameter values (Clark et al., 2011; Li et al., 2018), which can affect different physical parameterizations. Although model parameter tuning can help to alleviate this issue, determining optimal values is highly dependent on model structures and how model outputs are influenced by input parameters. Sensitivity analysis is commonly used to evaluate these input-output relationships and parameter importance, but traditional one-at-a-time (OAT) methods yield varying results depending on the interdependence of parameters, particularly within complex models, leading to issues of poor reproducibility and inability to generalize results. Consequently, more advanced variance-based techniques like the Sobol method, in the context of global sensitivity analysis (GSA, Saltelli et al., 2008), exhibit superior performance in such tasks, albeit being computationally intensive (Herman et al., 2013) and sometimes infeasible, especially when dealing with complex weather/climate models like the widely used Weather Research and Forecasting (WRF) model (Skamarock et al., 2021).

An alternative approach that avoids running numerous model realizations is the utilization of surrogate models or emulators (Queipo et al., 2005; O'Hagan, 2006; Forrester et al., 2008; Fernández-Godino et al., 2017; Kim and Boukouvala, 2020; Longo et al., 2020; Lamberti and Gorlé, 2021). A surrogate model/emulator is a simpler model trained using the input-output pairs of the original complex high-fidelity model, that can be used in substitution of it. The emulator makes the model process more computationally efficient in producing model realizations, while it still provides accurate predictions of the output variable. Machine Learning (ML) algorithms designed for regression tasks offer a computationally efficient means to build surrogate models to be used for sensitivity analysis (Engelbrecht et al., 1995; Shen et al., 2008; Muthukrishnan and Rohini, 2016; Antoniadis et al., 2021; Torres, 2021; Zouhri et al., 2022). Over time, a variety of algorithms have been tested in the literature and used in different fields.

These algorithms can also be used to extract feature importance, which has become a well-established methodology widely employed in different geoscience fields, such as landslide susceptibility (Yilmaz, 2010; Catani et al., 2013; Pradhan, 2013; Youssef et al., 2016; Kalantar et al., 2018; Lee et al., 2018; Zhou et al., 2018; Chen et al., 2020; Liu et al., 2021; Daviran et al., 2023; Elia et al., 2023), forest fire susceptibility (Oliveira et al., 2012; Bar Massada et al., 2013; Arpaci et al., 2014; Pourtaghi et al., 2016; Satir et al., 2016; Gigović et al., 2019), water quality assessment (Palani et al., 2008; Rodriguez-Galiano et al., 2014; Sarkar and Pandey, 2015; Haghiabi et al., 2018; Shah et al., 2021; Alqahtani et al., 2022; Trabelsi and Bel Hadj Ali, 2022), hydrological modelling ~~(Zhang et al., 2009)~~(Zhang et al., 2009; Yu et al., 2024), air quality assessment (Suárez Sánchez et al., 2011; Yu et al., 2016; Maleki et al., 2019; Sihag et al., 2019; Lei et al., 2023), groundwater mapping (Rahmati et al., 2016), agronomy (Kok et al., 2021; Sridhara et al., 2023; Wu et al., 2023), climatological applications (Wu et al., 2021; Dey et al., 2022), renewable energy (Wolff et al., 2017; Meenal et al., 2022), earthquake detection (Murti et al., 2022), and it also has significant relevance in civil engineering (Tian, 2013; Gholampour et al., 2017; Farooq et al., 2020; Salmasi et al., 2020), genetics (Sharma et al., 2014), biology (Cui and Wang, 2016), and medical research (Antonogeorgos et al., 2009; Maroco et al., 2011; Yang et al., 2022). ~~While~~

ML techniques have gained traction in weather and climate modeling and observations (Schultz et al., 2021; Schneider et al., 2022), particularly in parameter optimization tasks like calibration ~~(Bocquet et al., 2020; Bonavita and Laloyaux, 2020)~~ (Bocquet et al., 2020; Bonavita and Laloyaux, 2020; Williamson et al., 2013; Couvreux et al., 2021; Dagon et al., 2020; Watson-Parris et a
, spatial interpolation (Stein, 1999; Sekulić et al., 2020), downscaling (Fowler et al., 2007; Maraun and Widmann, 2018; Leinonen et al., 2021), parameterization substitution (Rasp et al., 2018; Han et al., 2020; Yuval and O'Gorman, 2020; Mooers et al., 2021; Grundner et al., 2022; Ross et al., 2023), and image-based classification (Chase et al., 2022, 2023).

Among the most relevant for the topic of the present work, Dagon et al. (2020) focused on building a surrogate model based on feed-forward artificial neural networks of a land surface model (CLM5) ensemble perturbed parameter which greatly improved the rapidity to generate predictions. Similarly, Cinquegrana et al. (2023) built a framework for optimizing physical parameters for the ICOsahedral Non-hydrostatic (ICON) Limited Area Mode at high resolution, aiming to reduce the discrepancy between observed and modeled meteorological variables using an Efficient Global Optimization algorithm relying on Gaussian-based surrogate models. Watson-Parris et al. (2021) introduced an open-source tool (ESEm) of surrogate modeling implemented to explore sensitivity in a cloud-resolving model and scenario uncertainty in the CMIP6 multi-model ensemble. Couvreux et al. (2021) used Gaussian process-based methods to calibrate parameters through the comparison of single-column simulations and reference large-eddy simulations over multiple boundary-layer cases.

Despite these recent advancements, the extraction of feature importance remains relatively uncommon in the meteorological/climate modeling literature~~(Ren et al., 2020; Baki et al., 2022)~~. Notable exceptions in recent times include Baki et al. (2022) , who employed global sensitivity analysis methods and a surrogate model based on Gaussian process regression, the study found that a subset of parameters significantly influenced the WRF simulations of tropical cyclones over the Bay of Bengal. Similarly, Fischer et al. (2024) used a Gaussian process regression-based surrogate model of the ICON model to quantify the uncertainty of simulations of the African monsoon through a global sensitivity analysis.

Many of the previous studies have proposed comparisons between several feature importance analysis algorithms. This is because the ability of these algorithms to best capture feature relevance is influenced by a variety of factors that can change with ~~the application~~applications, depending on the context under analysis, such as the degree of nonlinearity of the input-output relationships, the interaction degree between features, the dimensionality of the features, the size and quality of the data used for training, the shape and smoothness of the distribution of the training data, and ultimately the validity of each algorithm's assumptions. Often, the influence of one or more of the factors on the quality of the chosen method cannot be assessed in advance, leading to a trial-and-error procedure that would benefit from a multi-algorithm approach where the results of different methods can be compared. For this reason, the present work shares the same multi-method philosophy of many of the studies mentioned above, extracting the most popular algorithms available in the literature and combining them into a single flexible efficient framework for analysis.

The importance of sensitivity analysis in Earth science modeling is critical, not only for academic pursuits, but also for its practical implications for public safety and resource management. Currently, the diversity of research cultures across scientific disciplines, coupled with heterogeneous computational resources and varying degrees of familiarity with sensitivity analysis techniques, contributes to a predominant reliance on older, more familiar methods. This scenario prevails despite the increasing complexity of models, which would require more robust sensitivity analysis techniques. The field of meteorology currently exhibits a significant gap in the adoption of advanced sensitivity analysis methods, despite the chaotic nature of atmospheric dynamics and the interactions among numerous parameterizations in atmospheric models contribute to a high degree of sensitivity to input parameter variations, underscoring the need for robust uncertainty quantification to improve model reliability.

In light of the above considerations and to fill this gap, this paper proposes a new tool, the Machine Learning-based Automated Multi-method Parameter Sensitivity and Importance analysis Tool (ML-AMPSIT), that aims at providing a flexible and easy-to-use framework for performing sensitivity and importance analysis also for complex models. ML-AMPSIT applies a series of ML feature importance extraction algorithms to model parameters (using the widely-used WRF/Noah-MP coupled weather model as a case study), accommodating any user-specified model configuration. ML-AMPSIT represents a novel contribution to the field by providing a toolkit that integrates multiple machine learning algorithms for an improved sensitivity analysis. The included algorithms are among the most commonly used in literature, namely: LASSO, Support Vector Machine, Classification and Decision Trees, Random Forest, Extreme Gradient Boosting, Gaussian Process Regression and Bayesian Ridge Regression. These algorithms have been chosen for their simplicity and speed and to create an ensemble of state-of-the-art ML models each employing distinct methodologies, so as to improve the flexibility of the tool and its performance in different possible applications. This diversity allows for a robust method of self-validation or self-falsification of the results through comparative analysis, enhancing the reliability of the findings by ensuring that consistent results are not an artifact of a single modeling approach. While most of ~~them~~ these algorithms directly provide a measure of feature importance through the Python `scikit-learn` libraries, the last two methods are ~~probabilistic methods,~~ specifically used in this framework for a fast implementation of the Sobol method through the `SALib` Python library, leading to a computationally efficient way to obtain the Sobol sensitivity indices directly from the ML-inferred ~~relations~~ relation between input and output data. Our tool's objective is overall to assist users in evaluating parameter sensitivity and importance using computationally inexpensive and nonlinear interaction-aware approaches.

ML-AMPSIT guides the user through the different steps of the sensitivity and importance analysis, allowing, on the one hand, for a simplification and automatization of the process and, on the other hand, for extending the application of advanced sensitivity and importance analysis techniques to complex models, thanks to the use of computationally inexpensive and non-linear interaction-aware methods.

This paper is organized as follows: Section 2 outlines the methodology used to develop ML-AMPSIT, including a detailed description of the machine learning models integrated into the tool and the workflow for performing sensitivity and importance analysis. Section 3 presents the case study involving the coupled WRF/Noah-MP model to demonstrate the application of ML-AMPSIT. The results of the sensitivity analysis are discussed in Section 4, highlighting the effectiveness of different machine learning models in identifying the key parameters for the case study presented in this paper. Finally, Section 5 concludes the paper with a summary of the findings and some insights into potential future work to further enhance the capabilities of ML-AMPSIT.

2 Methods

In this section, we describe the methodological framework underlying this study. We begin with an overview of the ML-AMPSIT workflow, detailing the process from the selection of the input parameters to the sensitivity analysis phase. We then introduce the Sobol method, a variance-based technique used for GSA. Finally, we provide a description of the machine learning algorithms integrated into the tool, highlighting their main characteristics, how they are implemented and used in ML-AMPSIT and the rationale behind their selection.

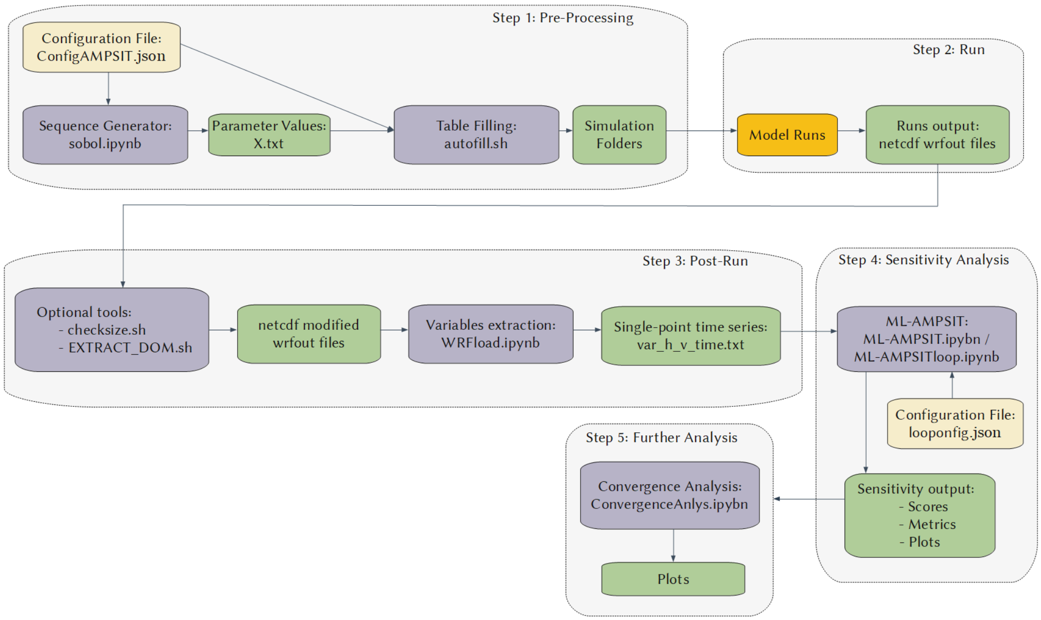

Figure 1. ML-AMPSIT workflow. The main code scripts are indicated with ~~the~~ blue boxes. The yellow boxes indicate the configuration files that need to be filled by the user, the green boxes refer to all the output files that eventually become inputs for other subsequent scripts, and the orange box indicates the generic model execution that varies depending on the model involved.

2.1 ML-AMPSIT workflow

The ML-AMPSIT workflow (Figure 1) can be divided into 5 main steps, each of which involves one or more Python/Bash-based scripts: the pre-processing phase, the model run phase, the post-processing phase, the sensitivity analysis phase, and an optional convergence analysis phase.

1. The selection of the input features is accomplished by specifying the parameter names within the configuration file *configAMPSIT.json*. The compiled configuration file related to the case study discussed in this paper is reported in the Appendix. There is no upper limit for the number of parameters that can be analyzed, but it is worth noting that the sensitivity analysis could converge significantly more slowly in high-dimensional (i.e., with more parameters) problems. Moreover, the scalability with the number of parameters can highly depend on the case study considered. The number of simulations to perform must also be specified through the configuration variable *totalsim*. To generate the values of the parameters to be tested, a Sobol sequence of the same length of *totalsim* is produced for each parameter from the pre-processing script *sobol.ipynb*. The Sobol sequence (Saltelli et al., 2010, Bratley and Fox, 1988) is a quasi-random low-discrepancy distribution designed to produce well-spaced points in the unit hypercube representing the parameter space. Unlike random sampling, each point in the sequence considers the positions of the previous points, resulting in

a more uniform filling of gaps, as shown in Figure 2. Consequently, a robust sequence is generated more efficiently compared to random sampling, requiring fewer points.

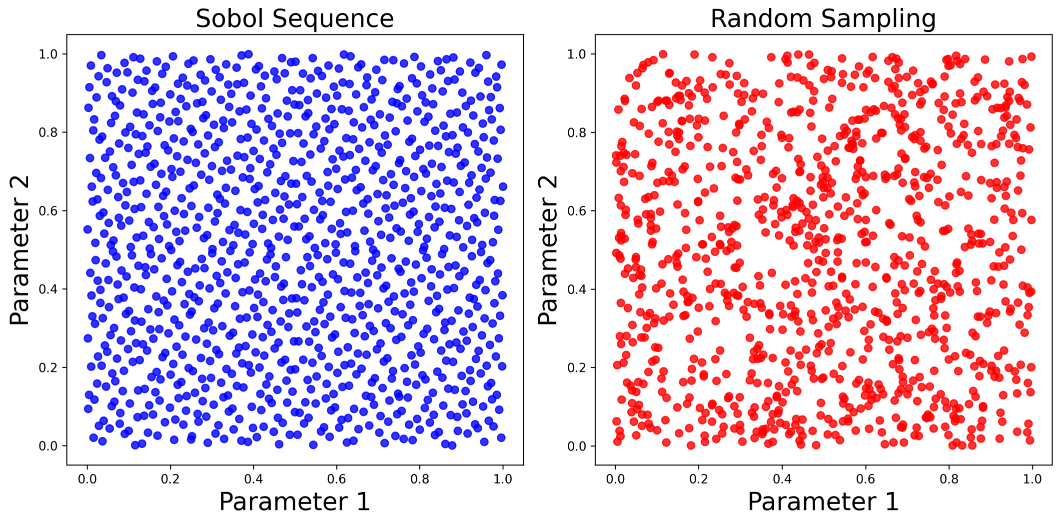

Figure 2. Demonstration of the differences between random sampling (red dots) and Sobol sampling (blue dots) in representing the parameter space. The Sobol sequence is able to more uniformly cover the parameter space, avoiding the presence of very close points, as it occurs in the random sequence.

Once the Sobol sequences are generated, a user-specified reference value and a maximum perturbation percentage need to be specified in *configAMPSIT.json*, which will be passed to the preprocessing script *autofill.sh*. These values are used to rescale the sequence values from the [0,1] range to the actual parameter range space. The output of the *sobol.ipynb* script is the file *X.txt*, containing a $m\ x\ n$ matrix, where m is the number of simulations and n the number of parameters tested.

Therefore, each row specifies a different set of parameter values that will be used in each particular model realization. Based on these data, the *autofill.sh* script creates multiple copies of the folder in which the model is run and then searches for each parameter name within the original model parameter look-up table in each newly created folder. The values of the parameters are then changed according to the *X.txt* file for each realization. Since this script edits the original model parameter look-up table, which is MPTABLE.TBL for the WRF/Noah-MP model in the case study, it is necessarily model-dependent and thus needs to be adapted if used with other models, to suitably modify the values of the tested parameters.

2. After all the simulation folders have been created, the original high-fidelity model can be run by the user as usual. It will be necessary to collect all the output files into one single folder, whose path must be specified in the configuration file, so that the post-processing script can find it.

3. Once the user has completed all the high-fidelity model runs, a post-processing script named *WRFload.ipynb* is provided to extract single-point time series for each output variable at specific coordinates in the simulation domain, as specified by the user in the configuration file. The resulting output data, which serves as input to the sensitivity analysis tool *ML-AMPSIT.~~ipybn~~ipynb*, consists of different files with the naming convention *var_h_v_time.txt*, where var is the variable name, h and v indicate the labels identifying the horizontal and vertical grid cell, while $time$ represents the simulated time. The script *WRFload.ipynb* specifically extracts variables from NetCDF files that follow a WRF-like format (a widely used format for weather/climate models). If the user's model output follows a different format, the script must be modified accordingly.

   Extracting single-point time series means that the tool has no information about spatial patterns, nor can it capture time patterns, because the sensitivity analysis is performed on each single point and time separately. Consequently, the importance of different parameters can be directly compared only for the specific point and time being analyzed. However, the observation of consistent relative importance between parameters across different points and times can confirm the statistical robustness of the results.

4. The sensitivity phase is performed by the main script of the ML-AMPSIT tool *ML-AMPSIT.~~ipybn~~ipynb*, which accomplishes a regression task based on different ML algorithms offered to the user. As mentioned in the Introduction, this multi-method approach is useful for comparing different results, since each algorithm is structurally different and could be more or less appropriate for the problem at hand. This ML-based ensemble philosophy is similar to an ensemble learning (EL) approach (Ren et al., 2016), which combines the predictions of multiple base models to improve the overall performance, but the present tool does not yet include an option to integrate the methods into a stacking, bagging, or boosting procedure, allowing for the user to choose any single or multiple methods independently. For each method offered by the present tool, the input and output data are split into training and testing sets in proportions of 70% and 30%, and each set is scaled separately to have zero mean and unit variance with respect to the ensemble. The training set is used to fit the model to the data, while the testing set is used to evaluate the model's ability to reproduce new data. This strategy is used to mitigate the risk of overfitting. The coefficient of determination (R^2), the mean squared error (MSE), and the mean absolute error (MAE) have been used as measures of goodness when comparing the predicted output against the actual "truth", i.e. the results of the original physical model simulations. Since all the variables are scaled before calculating these error metrics, MSE and MAE are not affected by the different scales of the variables. This allows for a fair and meaningful comparison of the model's performance across different variables. The coefficient of determination $R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$, is used as a measure of goodness of fit, where SS_{res} is the residual sum of squares and SS_{tot} is the total sum of squares. R^2 indicates how much variation in the target variable can be explained by the model's predictors. R^2 is typically a value between 0 and 1, where values closer to 1 indicate a better ability of the model to explain the variance in the data. Eventually, if the chosen model fits worse than the average value, then $\frac{SS_{res}}{SS_{tot}}$ can be greater than 1 and R^2 is negative. If the model has low values of MSE and MAE, but also low values of R^2, it might indicate that the relationship between the input data and the target variable cannot be properly explained in terms

of linear weights only. This is an indication of nonlinearity in the output response. In addition to the R^2 coefficient, the associated p-value is also computed and saved.

The script *SAML.~~ipybn~~ipynb* produces an interactive graphical user interface (GUI) built from the ipywidgets python library, which allows the user to specify which vertical level and surface point to consider in the analysis, the output variable for which to compute the sensitivity analysis, the number of simulations to consider, the algorithm to use, and the output time to plot for punctual evaluations. The flexibility of this GUI allows the user to quickly check the influence of the number of simulations on the robustness of the results and the performance of the different ML methods implemented.

Based on the selected options, unless the specified methods are Gaussian Process Regression or Bayesian Ridge Regression, the tool produces four plots: the upper two dedicated to the feature importance time series and the time-evolution of the metrics for the whole simulation duration, and the lower two showing the metrics and feature ranking specific to a particular time selected. Hence the user is provided with both the global result and the analysis related to a single output time. An example from the proposed case study will be provided in Section 3.3. If the specified method is either Gaussian Process Regression or Bayesian Ridge Regression, the features are ranked based on the total Sobol index, and the tool produces two additional plots, one showing the second-order Sobol interaction index between each couple of parameters, and the other showing the feature ranking based on the first-order Sobol index, which could potentially be different from the total ~~order~~ index based ranking if parameters' interactions are strong enough.

An alternative version of the main script is the ML-*AMPSITloop.ipynb* loop suite, which allows the user to automate multiple analyses by using the *loopconfig.json* configuration file, specifying all the combinations of settings to be explored, without the need to manually set each combination through the graphical interface, saving time for long in-depth analyses.

5. It is possible to further investigate the convergence of the obtained feature importance and metrics with the optional tool *convergenceAnlys.~~ipybn~~ipynb*. This script manipulates all the outputs of *ML-AMPSIT.~~ipybn~~ipynb* or *ML-AMPSITloop.~~ipybn~~ipynb* to visualize the sensitivity of the parameter scores to the number of simulations, the vertical grid point, the horizontal grid point, and the algorithm used. The plots produced by this tool, related to the case study presented in this paper, will be shown and commented on more in-depth in Section 4.2.

2.2 Sobol method

In classical sensitivity analysis, given a set of input parameters $\{X_1,..X_k\}$, the elementary effect of a single perturbation Δ in the input parameter X_i over the output $Y(X)$ is defined as ~~((Saltelli, 2007))~~(Saltelli, 2007):

$$EE_i = \frac{[Y(X_1,X_2,...,X_i+\Delta,...,X_k) - Y(X_1,X_2,...,X_i,...,X_k)]}{\Delta}, \tag{1}$$

The above definition assumes a linear relationship between the parameter and the output variable and it becomes ineffective in the presence of non-linearities or interactions between parameters. To achieve the highest level of generalization in sensitivity analysis, ~~it is required to evaluate~~ both the effect of the single input parameter and the additional effect of its interaction with other parameters must be evaluated.

A variance-based approach achieves this while also not relying on a linear assumption. The most well-established variance-based method is the Sobol method ~~((Saltelli and Sobol', 1995))~~(Saltelli and Sobol', 1995). With this approach, the variance $V(Y)$ is decomposed as:

$$V(Y) = \sum_i V_i + \sum_i \sum_{j>i} V_{ij} + \cdots + V_{12\ldots k}, \tag{2}$$

where V_i is the main effect variance, representing the contribution of the i-th input parameter to the output variance, V_{ij} is the second-order interaction effect variance, representing the combined contribution of the i-th and j-th input parameters to the output variance, and so on up to $V_{12..k}$, which represents the interaction effect variance of all k input parameters together.

Dividing by $V(Y)$, the Sobol indices are derived as $S_i = V_i/V(Y)$, $S_{ij} = V_{ij}/V(Y)$..., leading to:

$$1 = \sum_i S_i + \sum_i \sum_{j>i} S_{ij} + \cdots + S_{12\ldots k}. \tag{3}$$

The total effect, or total ~~order~~ index S_{Ti} is the total contribution to the output variation due to a specific factor X_i i.e. a specific parameter. Hence, for instance, for a set of three parameters X_1, X_2, X_3 the total effect index for the parameter X_1 is:

$$S_{T1} = S_1 + S_{12} + S_{123}. \tag{4}$$

Both first-order effects S_i and total effects S_{Ti} are important to assess the overall influence of an input parameter.

2.3 Implemented ML algorithms

2.3.1 LASSO

The Least Absolute Shrinkage and Selection Operator, (LASSO, Tibshirani, 1996) is an ML method used for feature selection and regression. It derives from the basic concept of curve fitting in the context of optimization, therefore it is one of the simplest algorithms among the most widely used. The goal of LASSO is to identify a subset of input features that are most predictive of the output variable, while also performing regularization to prevent overfitting.

In a classical regression problem, the goal is to find a function that maps the input features to the output variable. This is done by minimizing an objective function that takes into account the differences between the observations and the predictions, as a measure of how well the model fits the data. The minimization is performed during the training process to find the optimal values of model coefficients. In the LASSO algorithm, a penalty term (the "regularization") is added to the objective function, encouraging the model coefficients to be small. Specifically, the objective function for LASSO regression is the residual sum of squares (RSS), while the penalty term is based on the sum of the absolute values (L1 norm) of the coefficients, which promotes sparsity of the solution, and it can be adjusted to control the amount of shrinkage applied to the coefficients. As a result, LASSO can be particularly useful to identify the most important features by setting the coefficients of the less important features to zero.

The present tool uses LassoCV from the Python library ~~sklearn~~scikit-learn, which adds a cross-validation strategy to the standard LASSO algorithm. In cross-validation, data are divided into multiple subsets or "folds", and the model is trained and evaluated multiple times, each time using a different fold as the validation set and the rest of the data as the training set (e.g. as explained in ~~(Brunton and Kutz, 2019)~~Brunton and Kutz (2019)). By averaging the results of the multiple evaluations, cross-validation provides a more accurate estimate of the model's performance than a single evaluation on a single training-validation split.

2.3.2 Support Vector Machine

Support vector machine (SVM) is an ML method that can be used for regression as well as classification tasks. The original algorithm was proposed in 1963 (Vapnik, 1963), was refined multiple times (Boser et al., 1992; Cortes and Vapnik, 1995; Schölkopf and Smola, 2002), and is considered a core method in ML.

For regression tasks, SVM aims to find a hyperplane that approximates the underlying relationship between the input variables (i.e. the model parameters) and the continuous output values. In order to do this, SVM solves a convex optimization problem by minimizing a cost function that incorporates a margin of error and an L2-norm-based regularization term (ridge-type regularization). Unlike the LASSO-type regularization, the L2 regularization, also known as a ridge-type regularization, is a penalty based on the square of the coefficients. This penalty term is less strict compared to LASSO, because it helps to shrink the coefficient values towards zero without eliminating them completely. This can be preferred to a LASSO regularization in situations when there are many important parameters, since it avoids eliminating them in favor of the most important ones. The regularization term controls the trade-off between the complexity of the model and the amount of error allowed, and its strength is controlled by the hyperparameter "C". The optimization problem in the proposed implementation is solved by the default quadratic "liblinear" Python solver. SVM offers several advantages, including the ability to handle high-dimensional data and resistance to overfitting when properly regularized.

The tool presented in this paper implements a simple linear kernel, which maps the input data into a high-dimensional space using a linear function.

2.3.3 Classification and Regression Trees

The Classification and Regression Trees (CART, Breiman et al., 1984) algorithm is one of the most basic and straightforward decision tree methods, making it widely used and popular for various applications. The main idea behind CART is to recursively partition the input space into smaller regions based on the values of the input variables, with the goal of minimizing the impurity (or variance) within each resulting region. This partitioning process creates a binary tree structure called a "decision tree", where each internal node represents a splitting rule based on a selected input variable and a threshold value. The leaf nodes of the tree represent the final prediction or class assignment.

In the proposed tool, from the training dataset consisting of input-output pairs, the algorithm recursively selects the best splitting rule at each internal node based on the mean squared error for regression. The splitting process continues until a stopping criterion is met, such as a maximum tree depth or a minimum number of samples required to split a node. For regression tasks, the predicted output is the average value of the training samples within the leaf node.

2.3.4 Random Forest

Random Forest (RF, Breiman, 2001) is technically an ensemble learning method, that can be used both for classification and regression tasks. Differently from the CART algorithm, a multitude of decision trees are constructed in an RF, and the final prediction is based on the average of the predictions of all the trees. For this reason, RF usually performs significantly better than CART.

The advantage of using RF is that it can handle high-dimensional data and complex interactions between variables, making it a useful tool for sensitivity analysis. However, it can suffer from overfitting, where the model performs well on the training data but poorly on new data. To mitigate this, both cross-validation and Bayesian optimization are used to make the model more robust.

2.3.5 Extreme Gradient Boosting

The Extreme Gradient Boosting (XGboost, Chen and Guestrin, 2016) builds an ensemble of decision trees, similarly to RF, but each new tree is trained to correct the errors of the previous trees. It is hence a refined version of an RF and it is usually expected to perform better.

2.3.6 Gaussian Process Regression

Gaussian Process Regression (GPR), better known in geosciences as kriging (Stein, 1999) when applied to spatial interpolation, is a nonparametric algorithm for computing the probability density function of the regression curve, instead of a single fitting curve and can be used as a supervised ML technique (Rasmussen and Williams, 2005). GPR ~~is a non-parametric method, i.e. it does not make assumptions about the functional form of the relationship between the input and output variables. Instead, it models the relationship as a distribution over possible functions. GPR~~ assumes that the output values follow a Gaussian distribution with unknown mean μ and variance ~~Σ~~ σ which must be predicted, given a set of input-output pairs. To achieve this, GPR models the output values as a function of the input variables, where the function is assumed to be smooth and continuous. ~~The kernel implemented inside ML-AMPSIT, influencing~~ GPR is often described as a non-parametric method because it does not assume a specific functional form for the relationship between input and output variables. Instead, it models this relationship as a distribution over possible functions, allowing for flexibility in the shape of the regression curve. However, it is important to note that there are underlying assumptions about the functional form embedded in the chosen kernel. The kernel influences the shape and properties of the functions that the Gaussian ~~Process can learn,~~ process can learn. The kernel implemented inside ML-AMPSIT is the Radial Basis Function (RBF) kernel, which uses functions of the type $\sigma^2 \exp\left(-\frac{(t-t')^2}{2l^2}\right)$ where σ^2 is a variance, l is a length scale and t, t' represent pairs of values extracted from the training data. The choice of an RBF kernel is particularly advantageous due to its ability to model complex, non-linear relationships without imposing strong parametric constraints. The RBF kernel's smoothness assumption is well-suited for many real-world applications where the underlying function is expected to be continuous and differentiable. During the training phase, GPR estimates the parameters of this kernel function and calculates the covariance matrix between the input-output pairs. Using this covariance matrix and the training data, GPR then estimates the mean and variance of the distribution for each new input value through Bayesian inference. ~~In the proposed tool, a RBF kernel has been chosen due to its flexibility and absence of linearity assumptions.~~

~~Once built and tested against the original model outputs, the GPR surrogate model can be used to perform a GSA in substitution of the original model. By using a surrogate model, the computational cost of running the original model for a large number of input combinations is avoided. Instead, the surrogate model can be used to generate a large number of input combinations with significantly less computational time and evaluate their impact on the output. Over these samples, the Sobol sensitivity indices can be feasibly computed following for instance the definition proposed by Saltelli (Saltelli et al., 2008).~~

~~In the proposed tool, after the algorithm generates the optimal surrogate model, it uses the Python library SALib to compute the Sobol first-order index as a score for the sensitivity importance of each parameter.~~

2.3.7 Bayesian Ridge Regression

In Bayesian Ridge Regression (BRR), the hyperparameters of a classical ridge regression (i.e. a linear regression which implements a ridge regularization term) are associated with a priorly assumed probability distribution (also called hyperpriors), and tuned through training in a Bayesian inference approach (Box and Tiao, 1992). Defining both a prior distribution $p(H)$ for the model parameters H and a likelihood function $p(E|H)$ for the ingested data E, the BRR model computes the posterior distribution over functions $p(H|E)$ given the observed data through the use of Bayes' theorem $p(H|E) = \frac{p(E|H)\cdot p(H)}{p(E)}$, where $p(E) = \int p(E|H)\cdot p(H)\,dH$ is the marginal likelihood. Once the posterior distribution is obtained, the model is used to make predictions for unseen data points. These predictions come with uncertainty estimates, which are derived from the posterior distribution. BRR, as in the case of the SVM algorithm, employs an L2 regularization, hence it spreads the coefficient values more evenly, stabilizing the model and preventing overly large coefficient estimates.

~~The same procedure used for the GPR algorithm to leverage the probabilistic output for deriving feature importancecoefficients is also implemented here to compute the Sobol first-order sensitivity index. Therefore,~~

2.3.8 Feature importance computation

Each of the algorithms implemented in this study provides a method for calculating feature importance, albeit through different approaches. In principle, a single sensitivity method could be used to evaluate feature importance across all algorithms. However, some algorithms have built-in methods specifically designed to align with their inherent characteristics.

– Fitting Methods: LASSO and SVM derive feature importance from the model coefficients. In these linear models, the magnitude of the coefficients indicates the strength and direction of the relationship between each feature and the target variable. Specifically, in the `scikit-learn` library, this can be accessed through the `best_estimator_.coef_` attribute. Larger absolute values of these coefficients indicate greater importance.

– Tree-based algorithms: for CART, RF, and XGboost, feature importance is assessed using the Mean Decrease in Impurity (MDI) method. This method quantifies the contribution of each feature to the ~~user can~~ overall prediction accuracy by measuring how much each feature decreases the impurity of the splits in which it is involved. For RF and XGboost, the final value is obtained by averaging over all the trees in the ensemble. In `scikit-learn`, these contributions are accessible through the `feature_importances_` attribute. The MDI method is particularly effective because it directly measures the impact of each feature on the model's decision process, providing a clear indication of feature importance.

– Probabilistic methods: GPR and BRR do not have a built-in mechanism for directly assessing feature importance. Therefore, in this work, the Sobol method was used to infer feature importance. Once built and tested against the original model outputs, the GPR and BRR surrogate models can be used to perform a GSA in substitution of the original model. By using a surrogate model, the computational cost of running the original model for a large number of input combinations is avoided. Instead, the surrogate model can be used to generate a large number of input combinations with significantly less computational time and evaluate their impact on the output. Over these samples, in ML-AMPSIT the Sobol sensitivity indices are computed following the definition proposed by Saltelli et al. (2008). The user can then compare the Sobol indices evaluated with both ~~these two methods~~GPR and BRR, providing information on their robustness and reliability. In the proposed tool, after the algorithm generates the optimal surrogate model, it uses the Python library `SALib` to compute the Sobol first-order index as a score for the sensitivity importance of each parameter. The Sobol total index and Sobol second-order interaction term are available for users who wish to examine the presence of strong parameter interactions.

Despite the differences in the feature importance calculation approaches of the different algorithms, each method is applied to standardized, non-dimensional data and each feature importance set is scaled between [0,1]. This ensures that feature importance scores are comparable across models. The primary objective of all these methods is to quantify the sensitivity of the model output to changes in the input features. Consequently, the feature importance scores obtained from these different methods provide a well-posed comparison of parameter sensitivities. By evaluating and comparing these scores, it is possible to gain a comprehensive understanding of the relative importance of each feature across different modeling approaches, which increases the robustness of the results.

2.3.9 Hyperparameter Tuning

In the proposed tool, the hyperparameters of each implemented algorithm are tuned based on a cross-validation score obtained through Bayesian optimization. Bayesian optimization is an iterative process that seeks to explore the hyperparameter space while also exploiting regions of the space that are expected to yield good performance. At each iteration, the method proposes a new set of hyperparameters based on a probabilistic model of the function behavior and then evaluates the function at that point. The results of the evaluation are used to update the probabilistic model, which is then used to propose a new set of hyperparameters for the next iteration.

Table 1. Summary table of the ML models' characteristics implemented in ML-AMPSIT

Model Name	**Regression Type**	**Feature Importance Method**	**Additional Characteristics**
LASSO	Linear	Model Coefficients	Applies L1 regularization to encourage sparsity in coefficients
SVM	Linear	Model Coefficients	Uses L2 regularization; implemented with a linear kernel
CART	Nonlinear	Mean Decrease in Impurity (MDI)	Constructs binary trees for decision-making based on feature values
RF	Nonlinear	Mean Decrease in Impurity (MDI)	Ensemble method that builds multiple decision trees
XGBoost	Nonlinear	Mean Decrease in Impurity (MDI)	Gradient boosting technique, correcting errors of previous trees
GPR	Nonlinear	Sobol Indices (external method)	Models output as a Gaussian distribution; uses Radial Basis Function kernel
BRR	Linear	Sobol Indices (external method)	Incorporates Bayesian inference with Ridge (L2) regularization

To conclude this methods section, Table 1 provides a summary of the main characteristics of the machine learning models used in this study. This table outlines the regression type of each model (linear or nonlinear), the associated feature importance methods used, and other relevant characteristics such as regularization techniques and kernel functions.

3 Case study

The proposed case study, adopted to highlight the functionalities of ML-AMPSIT, is based on idealized coupled WRF/Noah-MP simulations of a sea breeze circulation over a flat three-dimensional domain. The objective of this case study is to evaluate the impact of a prescribed set of Noah-MP parameters on the development of the thermally-driven wind.

3.1 The WRF/Noah-MP model

The Weather Research and Forecasting (WRF) model is a widely-used state-of-the-art mesoscale numerical weather prediction model for atmospheric research and operational forecasting applications, which is supported by the NSF National Center for Atmospheric Research (NCAR), with more than 50000 registered users from more than 160 countries (Skamarock et al., 2021). It offers a wide range of customization options consisting of dedicated modules and physics schemes to meet the state of the art in atmospheric science and adapt to a wide variety of scenarios. The dynamical core used for this case study is the Advanced Research WRF (ARW), which uses a third-order Runge-Kutta scheme for time integration with a time-split method for solving acoustic modes (Wicker and Skamarock, 2002) and an Arakawa-C grid staggering for spatial discretization. In the case study presented here to illustrate the functionalities of ML-AMPSIT, WRF is coupled with the Noah with Multi-Physics options (Noah-MP, He and Ek, 2023 Niu et al., 2011, Yang et al., 2011) Land Surface Model (LSM). Noah-MP is one of the most used LSMs available in WRF to calculate surface-atmosphere exchanges and interactions. It is an augmented version of the Noah land surface model (Ek et al., 2003), that allows the usage of different physical schemes and multi-parametrization options, reaching a total number of 4,584 possible combinations (https://www.jsg.utexas.edu/noah-mp/).

3.2 Model setup

Simulations are performed using one domain with 201 x 201 cells in the horizontal plane with a grid spacing of 3 km. 65 vertical levels are used, transitioning from a vertical resolution of 7 m close to the surface up to 600 m at the top of the domain, which is placed at 16 km above sea level. The domain is completely flat and subdivided into two equally-sized rectangular sub-regions of land and water, with the interface line oriented along the west-east direction. The aim is to simulate the diurnal cycle of a sea/land breeze. Boundary conditions are set as open at all the boundaries.

The atmospheric initial potential temperature profile is set using the following expression, representative of a stable atmosphere:

$$\theta(z) = \theta_s + \Gamma z + \Delta\theta(1 - e^{-\beta z}), \tag{5}$$

~~with~~ where the surface temperature $\theta_s = 280$ K, $\Gamma = 3.2$ K ~~Km~~km^{-1}, $\Delta\theta = 5$ K, and $\beta = 0.002$ m^{-1}. The atmosphere is initially at rest and the relative humidity is set constant over all the domain and equal to 30%. The sea temperature is set as 293 K.

The idealized simulations start at 13:00 UTC, 19 March, are centered at 47°N, 11°E, and last for 35 hours, thus with a solar radiation cycle representative of the equinoxes at mid-latitudes. The first eleven simulation hours are not analyzed and considered as spin-up period. Therefore, analyses concentrate on a full diurnal cycle, from 00:00 UTC 20 March to 00:00 UTC 21 March. Model output is saved every hour.

The physical parameterization schemes selected for the present work are the Rapid Radiative Transfer Model (RRTM) for longwave radiation (Mlawer et al., 1997), the Dudhia scheme (Dudhia, 1989) for shortwave radiation and the YSU scheme (Hong et al., 2006) as planetary boundary layer (PBL) parameterization, coupled to the MM5 similarity scheme for the surface layer. Since this idealized study aims at reproducing a sea/land circulation, which best develops under completely clear sky, the microphysics parameterization is switched off, as well as the convective scheme, since convection is explicitly resolved at the resolution used.

As said above, the Noah-MP model is used ~~for evaluating~~ to evaluate land-atmosphere exchange. In Noah-MP, the canopy radiative transfer scheme used is the "modified two-stream" (Niu and Yang, 2004), one of the most used, which aggregates cloudy leaves into evenly distributed tree crowns with gaps. The gaps are computed according to the specified vegetation fraction. The Ball-Berry scheme, the most common choice in literature, is used for the stomatal resistance computation, with the Noah-type soil moisture factor (Schaake et al., 1996), while the ~~default TOPMODEL~~ surface runoff parameterization TOPMODEL (Niu et al., 2007) with groundwater option is used for runoff and groundwater processes~~(Niu and Yang, 2007)~~. The surface resistance to evaporation and sublimation processes is set following Sakaguchi and Zeng (2009). The surface-layer drag coefficient, used to compute heat and momentum exchange coefficients, is calculated with the traditional Monin-Obukhov-based parameterization. In this work, the dynamic vegetation option is not activated, with the prescription of a fixed vegetation fraction of 60% to consider a reasonably realistic percentage, while the monthly satellite-based climatological leaf area index is read from the MPTABLE.TBL file, which contains all the parameter values. Finally, crop and irrigation options are deactivated. It is important to underline that water physical properties are not varied in the sensitivity simulations, but changes in atmospheric variables are expected also over water, due to the indirect effects of the variations in the surface parameters over land.

In order to simplify this demonstrative case study, only six of the surface parameters defined in the look-up table MPTABLE.TBL, from which WRF reads and sets the surface values accordingly, are considered. In particular, the Noah-MP reference vegetation type adopted over land is grassland (vegtype=10 of the IGBP-Modified MODIS classification) and the parameters for which the sensitivity and relative importance is evaluated are the characteristic leaf dimension (DLEAF), the height of the vegetative canopy top (HVT), the momentum roughness length (Z0MVT), the near-infrared leaf reflectance (~~RHOL-NIR~~RHOL_NIR), the empirical canopy wind parameter (CWPVT), and the leaf area index referred to the month of March (hereafter LAI_MAR). The choice of these parameters is based on their importance in other sensitivity studies reported in the literature (Mendoza et al., 2015; Cuntz et al., 2016; Arsenault et al., 2018). The final perturbed model parameter ensemble contains 100 samples, each with different parameter values based on the associated Sobol sequences. The input ensemble is generated by perturbing the parameters by up to 50% of their reference value in the look-up table MPTABLE.TBL. It should be clear that the results of a sensitivity analysis, regardless of the approach chosen, always depend on the range of exploration of the parameters, and that their transferability to arbitrary ranges of values is not guaranteed if the true sensitivity of the parameters in unexplored ranges is not known a priori. The perturbation percentage in this work has been chosen to avoid unphysical values, but it must be noted that the aim of the present work is to introduce and test ML-AMPSIT functionalities in a simplified case study, while a more detailed analysis would require more attention to the choice of the parameter space.

The output variable for which sensitivity is evaluated is the south-north horizontal component of the wind v in the lowest 10 vertical levels at two different locations in the domain, one over land and one over water. These two locations are chosen to evaluate the effects of varying land parameters over two completely different surfaces and to assess how changes in land properties can influence atmospheric fields also over water. The locations are also strategically chosen near the interface between the land and water regions to better capture the dynamics of the sea/land breeze circulation, which is expected to be most pronounced near this boundary. For both locations, the average output from three adjacent cells in the south-north direction is analyzed, in order to increase the representativeness of the results. The central points of these two locations are x=100, y=95 and x=100, y=105 for water and land respectively, i.e., in the central cell in the west-east direction, and 5 grid points to the north and to the south of the land-water interface. Figure 3 showcases the domain setup.

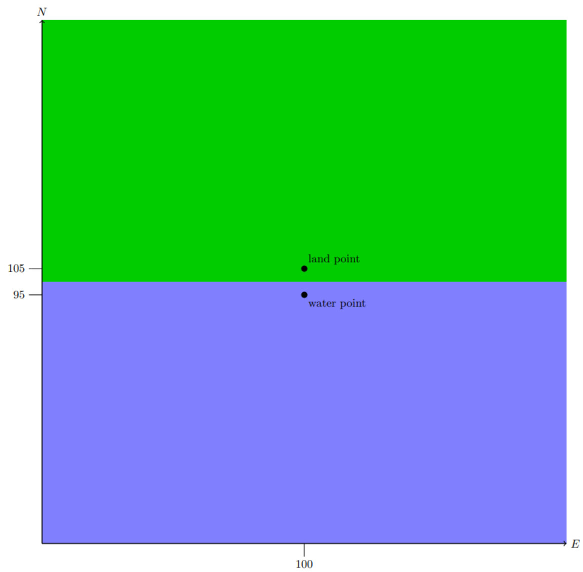

Figure 3. Domain configuration for the idealized simulation study. The grid consists of 201 x 201 cells with a 3 km spacing, divided equally into land (upper region, shaded in red) and water (lower region, shaded in blue). The land and water points for sensitivity analysis are marked, positioned at (x=100, y=95) for water and (x=100, y=105) for land. The interface line runs along the west-east direction, reflecting the setup for simulating a diurnal cycle of a sea/land breeze developing in the north-south direction.

4 Results

4.1 Sea breeze ensemble

Before analyzing the ML-AMPSIT results, this section presents an overview of the output of the WRF/Noah-MP simulations, focusing on the horizontal south-north wind component v at the two locations selected for the application of ML-AMPSIT. Figure 4 shows the ensemble time series of v during the entire period analyzed at the first vertical level. The daily cycle of the sea breeze is evident at both locations, as the velocity changes sign according to the radiation pattern and the varying horizontal temperature and pressure gradient, i.e. v is negative (northerly, from land to sea) during the night and positive (southerly, from sea to land) during the day. ~~The~~ It is worth noting that, even if only land parameters have been considered in this work, the spread of the ensemble tends to be larger over water than over land, especially before sunrise~~, indicating that the variation of the input parameters has a larger effect on v over water~~. Indeed, changes in land parameters affect the thermal contrasts between land and water, and thus the characteristics of the sea and land breeze, including their timing and strength. This highlights that changes in surface parameters can influence atmospheric variables not only locally, especially when they affect the development of thermally-driven circulations.

Figure 5 shows the ensembles of the vertical profiles of v in the lowest 200 m, containing the first 10 vertical levels, at three different times, 7:00 UTC, 13:00 UTC, and 19:00 UTC, which are representative respectively of the maximum intensity of the northerly land breeze, the morning transition between northerly and southerly wind, and the maximum intensity of the southerly sea breeze at the water point. It can be noted that the northerly land breeze (Fig. 5a,b) is shallower than the southerly sea breeze (Fig. 5e,f). Moreover, the comparison between Fig. 5e and Fig. 5f highlights the stronger effect of friction over land, with a more pronounced decrease of the wind speed close to the surface.

The ensemble variance is small near the ground over land and increases with ~~elevation~~ height at 07:00 UTC and 19:00 UTC, ~~while over water~~ which is expected for similar reasons as described above concerning the difference between the land and water ensembles. Over the water point, the spread is larger and more uniform along the entire vertical profile, especially at 13:00 UTC and 19:00 UTC. During the transition (Fig. 5c and Fig. 5d), the ensemble spread is very small over land, with all the simulations showing very small v values along the entire vertical profile, while a large spread is observed over water, suggesting that the variations in the surface parameters investigated significantly influence the timing of the transition from land to sea breeze over water, although preserving the shape of the vertical profile.

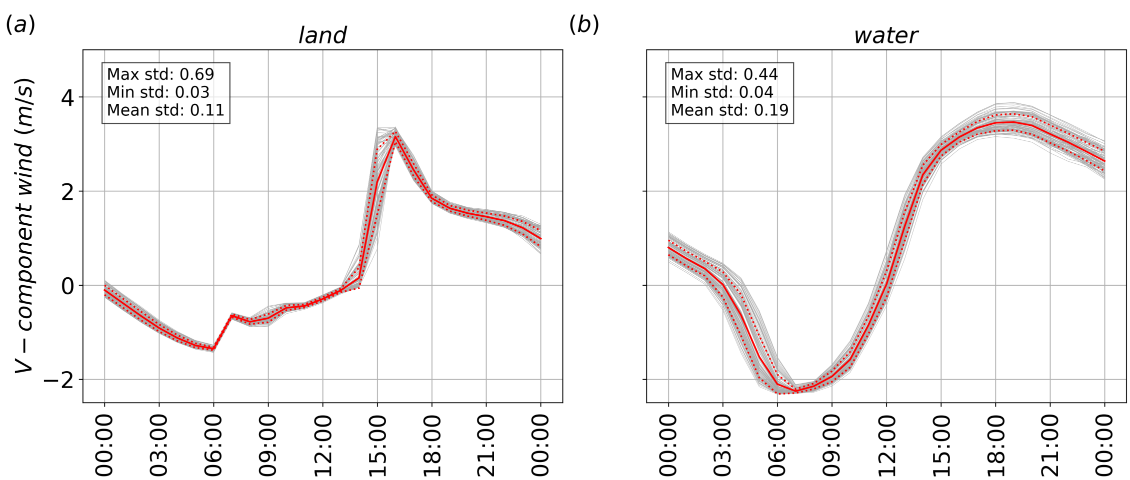

Figure 4. Time series of the south-north wind component at the first vertical level. The solid and dashed red lines represent the ensemble mean and standard deviation, respectively, while the gray lines represent the output of the single simulations. The values of the maximum and minimum hourly standard deviations (std) are also reported.

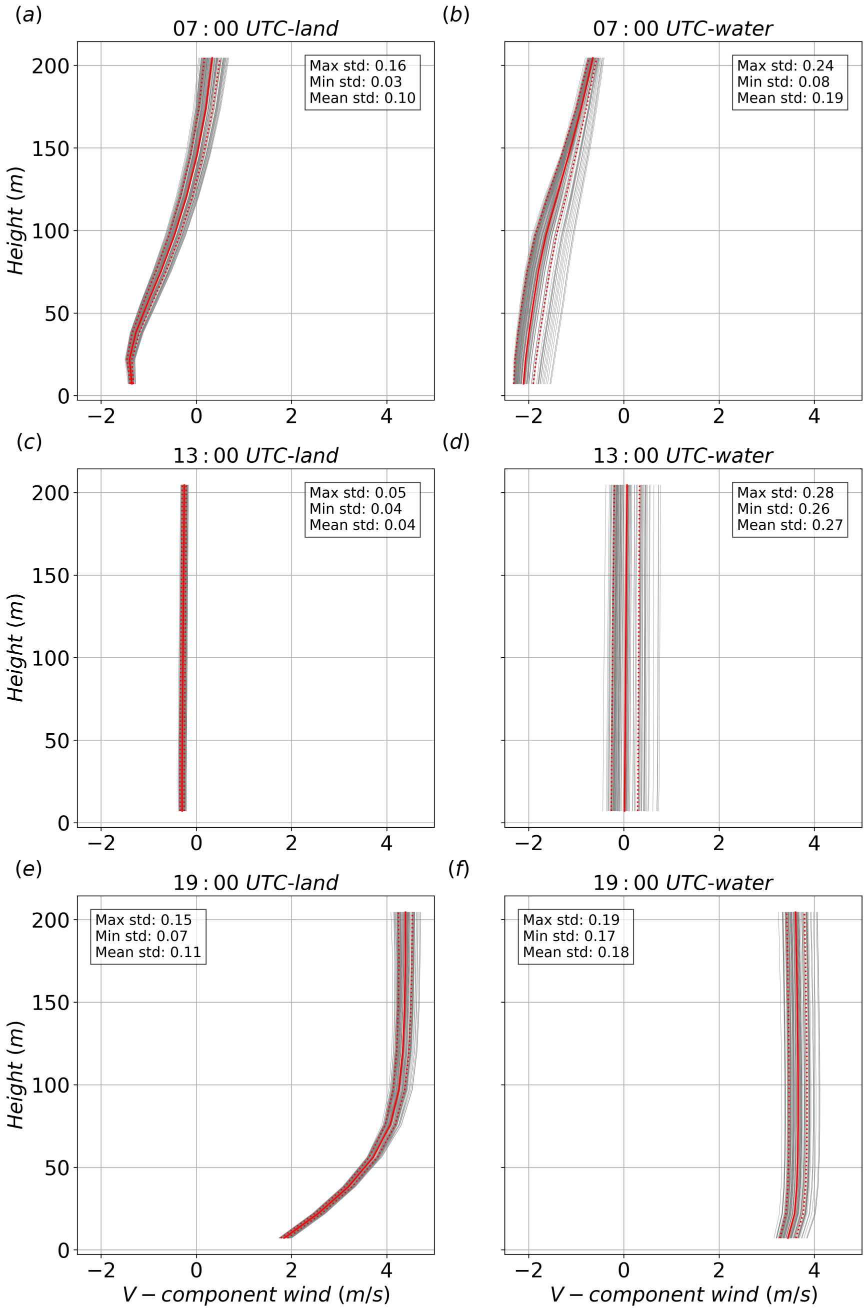

Figure 5. Vertical profiles of the south-north wind component in the lowest 200 m above the surface, at a) 07:00, b) 13:00, and c) 19:00 UTC. The solid and dashed red lines represent the ensemble mean and standard deviation, respectively, while the gray lines represent the output of the single simulations. The values of the maximum and minimum hourly standard deviations (std) are also reported.

4.2 ML-AMPSIT output

After all the steps discussed in Section 2.1, the basic output that the ML-AMPSIT tool offers to the user is a composition of four plots similar to those reported in Figure 6, which refers to the results obtained for v at the third vertical level, over the water point (for this example, no averaging over multiple points was performed and all the simulation hours are shown, in order to present the default output of the tool) obtained with the LASSO algorithm. In particular, panel a) shows the importance time series related to each of the 6 selected parameters, panel b) shows the time evolution of the metrics, underlining how the correlation score and the errors eventually change over time, while panels c) and d) are specific to a single hour selected by the user, showing respectively the goodness of fit and the ranking of the importance of the features for that hour. In the following sections, the outputs from ML-AMPSIT will be aggregated to perform convergence analysis and methods comparison.

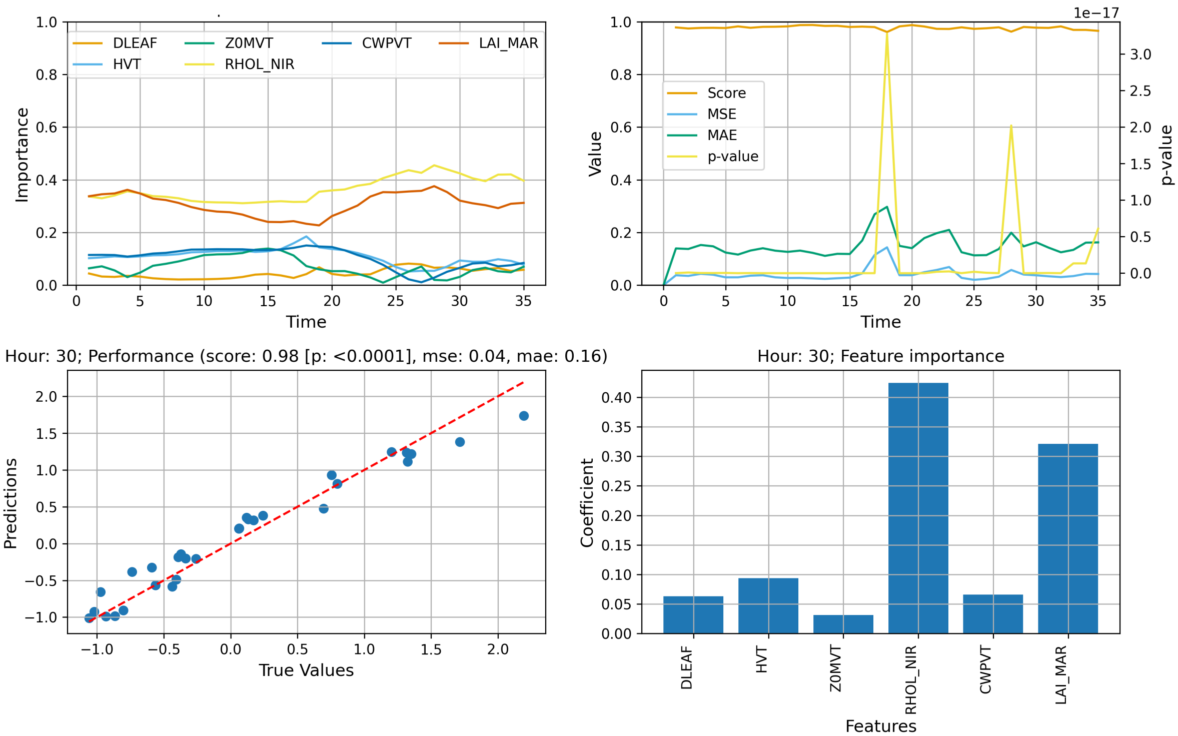

Figure 6. Example of the ML-AMPSIT output for a LASSO regression of 100 simulations with a duration of 36 hours (spin-up phase included), focusing on v at the third vertical level above the water point and displaying the metrics for the 30th simulated hour: a) importance time series of the 6 parameters; b) time evolution of the metrics MSE, MAE and p-value; c) quality of the test phase associated to the selected hour with the corresponding metrics; d) ranking of the importance of the features for the selected hour.

4.3 Convergence analysis

Figure 7 and 8 show the convergence of the MSE and of the feature importances computed by each method as a function of the number of simulations considered, i.e. the number of input-output relations used for training the algorithms. The analysis of the convergence is important because it indicates when the regression tasks reach a stable state and additional simulations do not significantly alter the results. When convergence is reached, it can be assumed that the obtained feature importance provides a reliable representation of the underlying relationships in the system and that a sufficient number of simulations are performed to capture the essential characteristics of the system under investigation.

For the sake of brevity, ~~here~~ only results over the land region at the first vertical level are shown here, considering results at 13:00 UTC for the ~~metrics convergence and at 8~~ feature importance and at 4 different times for the ~~feature importance~~MSE convergence. However, the considerations reported below can be generalized to other points/times, since the methods maintain similar speed of convergence during the entire run in every analyzed point.

Despite each method showing some differences, especially in the oscillations around the convergence values, four out of seven of the proposed methods are able to reach a reasonably stable result after approximately 20 simulations. BRR~~and GPR~~, GPR, LASSO and SVM are the fastest and most stable algorithms to reach convergence~~followed by LASSO and SVM~~. On the other hand, the decision-tree-based methods have significant oscillations even after 20 realizations and the metrics show that they are less consistent than the other methods. However, the results highlight that all the methods propose a stable and consistent ranking of the parameters' importance after 80 simulations and, in most cases, even with a much lower number of simulations.

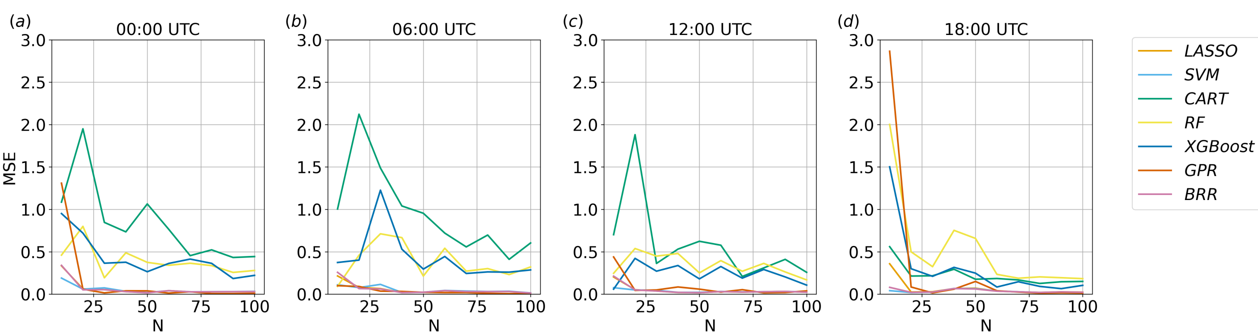

Figure 7. Convergence of the MSE with the number of realizations for each method implemented, considering v over land, at the first vertical level, at ~~8~~ 4 different times.

4.4 Parameter importance analysis

Figures 9 and 10 show respectively the time series of the performance metrics for the south-north wind component at the lowest vertical level over the land and water regions, while Figures 11 and 12 show, for the same variable and regions, the time series of the feature importance for each of the proposed methods. It is important to underline that comparing the results of each

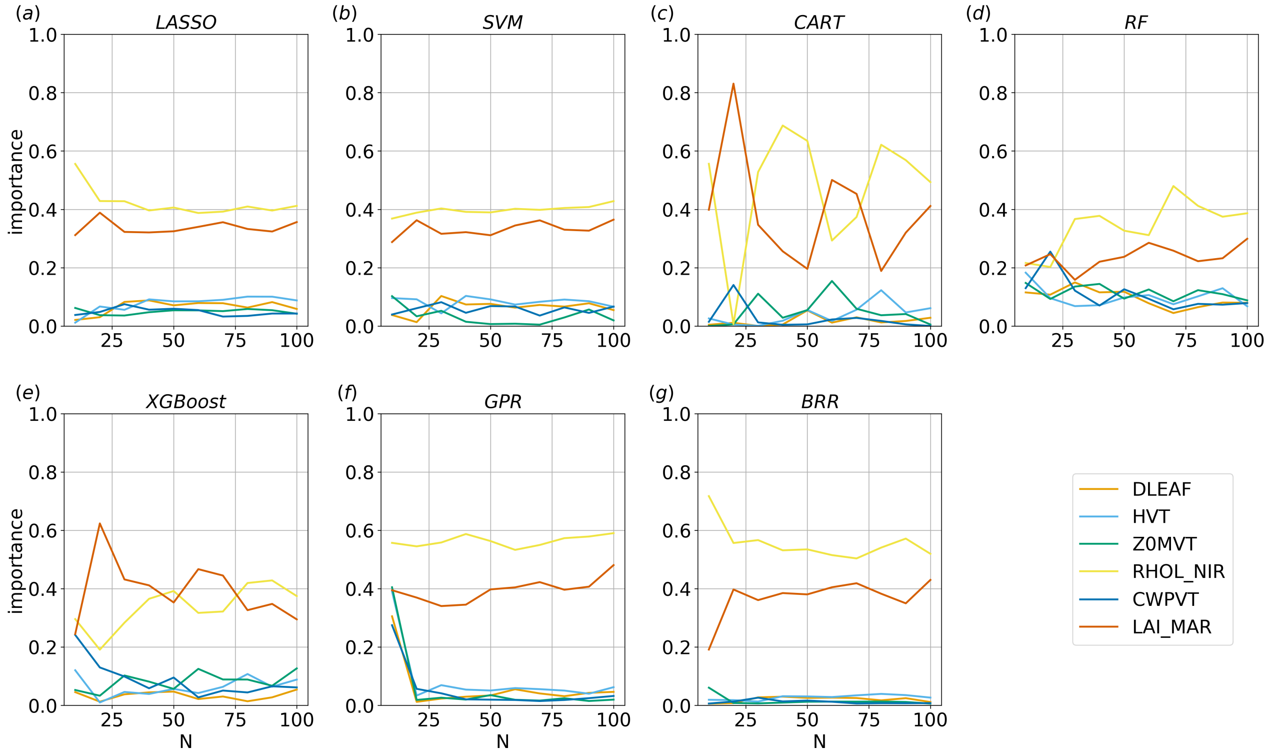

Figure 8. Convergence of the feature importance with the number of realizations, for each method implemented, considering v over land, at the first vertical level, at 13:00 UTC.

surrogate model is the core of ML-AMPSIT's robustness strategy. The agreement between the different models strengthens the reliability of the results and provides a form of self-validation.

In this example, GPR~~shows the best metrics, closely followed by LASSO, SVM and BRR~~, BRR, LASSO and SVM show the best metrics, suggesting that in the proposed case study there is no relevant difference between non-linearity-aware approaches and linear approaches, as they both correctly capture the relation between the tested parameters and the south-north wind component. These algorithms show very stable results, with slightly worse performance metrics ~~only in correspondence with the peak of the sea breeze around 16~~occurring around 13:00 UTC over land ~~and with the nocturnal peak of the northerly wind around 07~~(Figure 4a) and around 06:00 UTC over water (Figure 4b). These times correspond to sudden changes in the ensemble spread, but the observed degradation in performance metrics is likely due to the contrasting behavior of the ensemble members rather than the time variation itself, which ML-AMPSIT cannot be aware of by design. Around these timestamps, individual ensemble members exhibit divergent behavior, with some increasing and others decreasing, which can create challenges for the regression model to accurately capture the dynamics. The three decision-tree-based methods present a more irregular behavior of the performance metrics, with higher errors and lower correlations. In particular, CART presents the worst performance

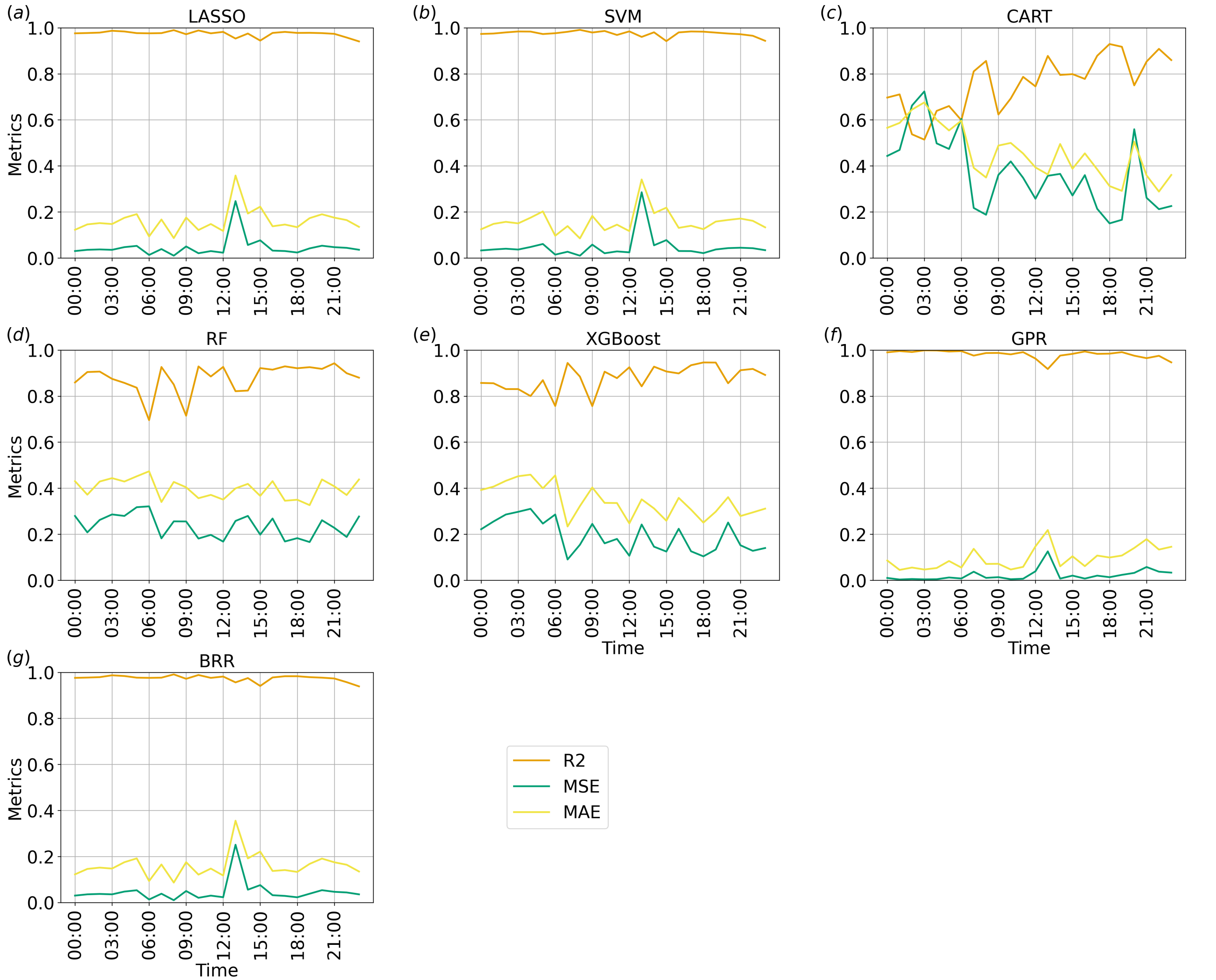

Figure 9. Time series of the performance metrics for each method implemented, considering v over land at the first vertical level: blue, green, red and orange lines represent respectively R^2, MSE ~~, MAE~~ and ~~p-value~~MAE.

metrics for this case study. The poorer metrics compared to more refined methods such as RF or XGboost are expected, since CART does not compute an ensemble of decision trees and does not consider the errors of the previous branches.

~~For both regions~~As shown in Figures 11 and 12, all methods agree very well for both regions on the ranking and overall magnitude ratios of the feature importance, individuating similar patterns, with only minor differences, also considering the methods showing worse performance metrics (cf. Fig. 9).

Figure 11 highlights cyclic trends of the parameters' importance over the land region, likely induced by the cycle of the diurnal thermally-driven circulation. In particular, Z0MVT and ~~RHOL-NIR~~ RHOL_NIR alternate as the most important pa-

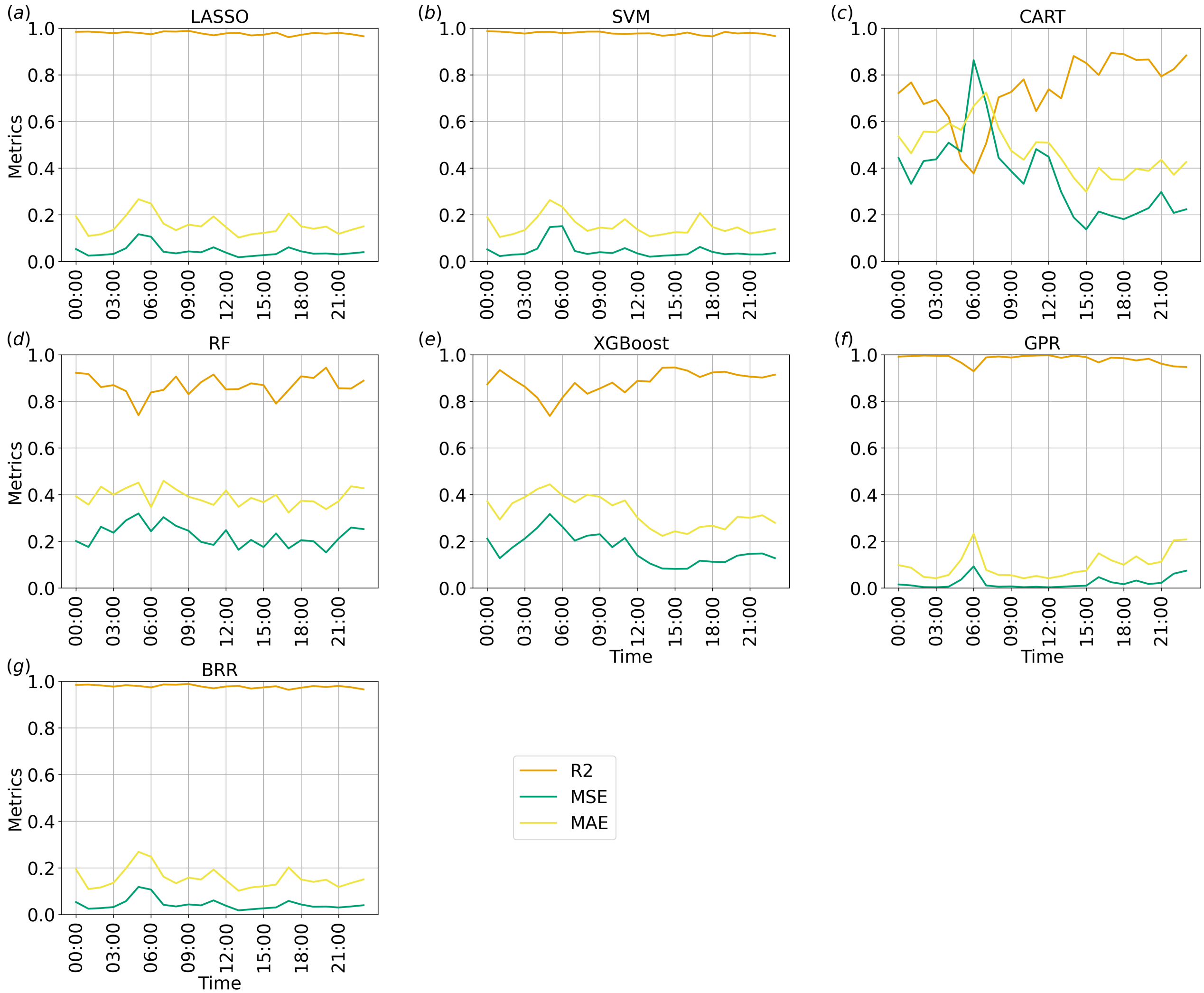

Figure 10. Time series of the performance metrics for each method implemented, considering v over water at the first vertical level: blue, green, red and orange lines represent respectively R^2, MSE ~~, MAE~~ and ~~p-value~~MAE.

rameters, with ~~RHOL-NIR dominating during the entire day, with the exception of the hours~~ RHOL_NIR dominating for most of the day, whereas Z0MVT becomes more important close to sunrise and sunset~~, when ZOMVT becomes more important~~. The short time windows in which Z0MVT appears as the dominant parameter correspond to the phases in which the vertical wind profile over land showcases the most pronounced shear in the lowest layers, as shown in Figure 5a,e. This seems to indicate a stronger role of ~~the surface friction~~ surface friction in dictating ensemble variability when stronger winds are present (~~ZOMVT directly influences the~~ Z0MVT directly influences surface friction)~~, coherently with the wind speed profiles shown in Figure 5a,e~~. LAI_MAR shows an importance almost comparable to ~~RHOL-NIR~~ RHOL_NIR during the day, especially with LASSO,

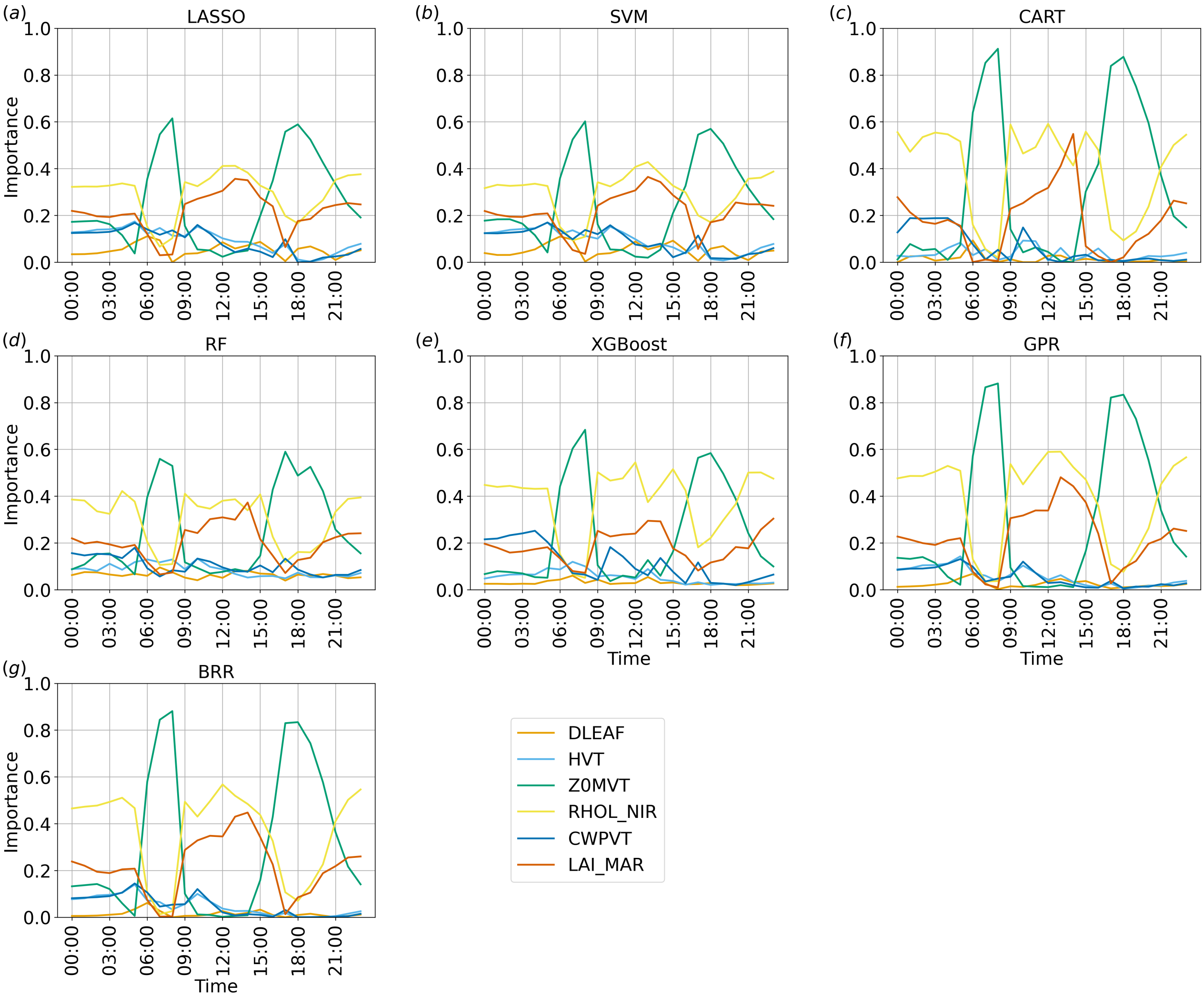

Figure 11. Time series of the importance of the parameters considered for each method implemented, considering v over land at the first vertical level.

SVM and RF, ~~while~~ whereas its importance is lower during the night, especially in the algorithms that tend to separate more the relevant and non-relevant parameters (i.e., GPR, BRR and decision-tree-based methods). The other parameters seem to be more relevant at night, with the exception of DLEAF, which is always non-relevant for each method implemented.

~~Over water, the~~ The results are more uniform over water than over land, and the ~~parameters' ranking~~ ranking of the parameters does not show significant variations during the whole day. In particular, the dominant parameters are ~~RHOL-NIR~~ RHOL_NIR and LAI_MAR, with Z0MVT always showing low importance values. Since the sea breeze is driven by thermal contrasts, it is expected that the parameters mainly affecting temperature, such as the reflectivity and the leaf area index, are also

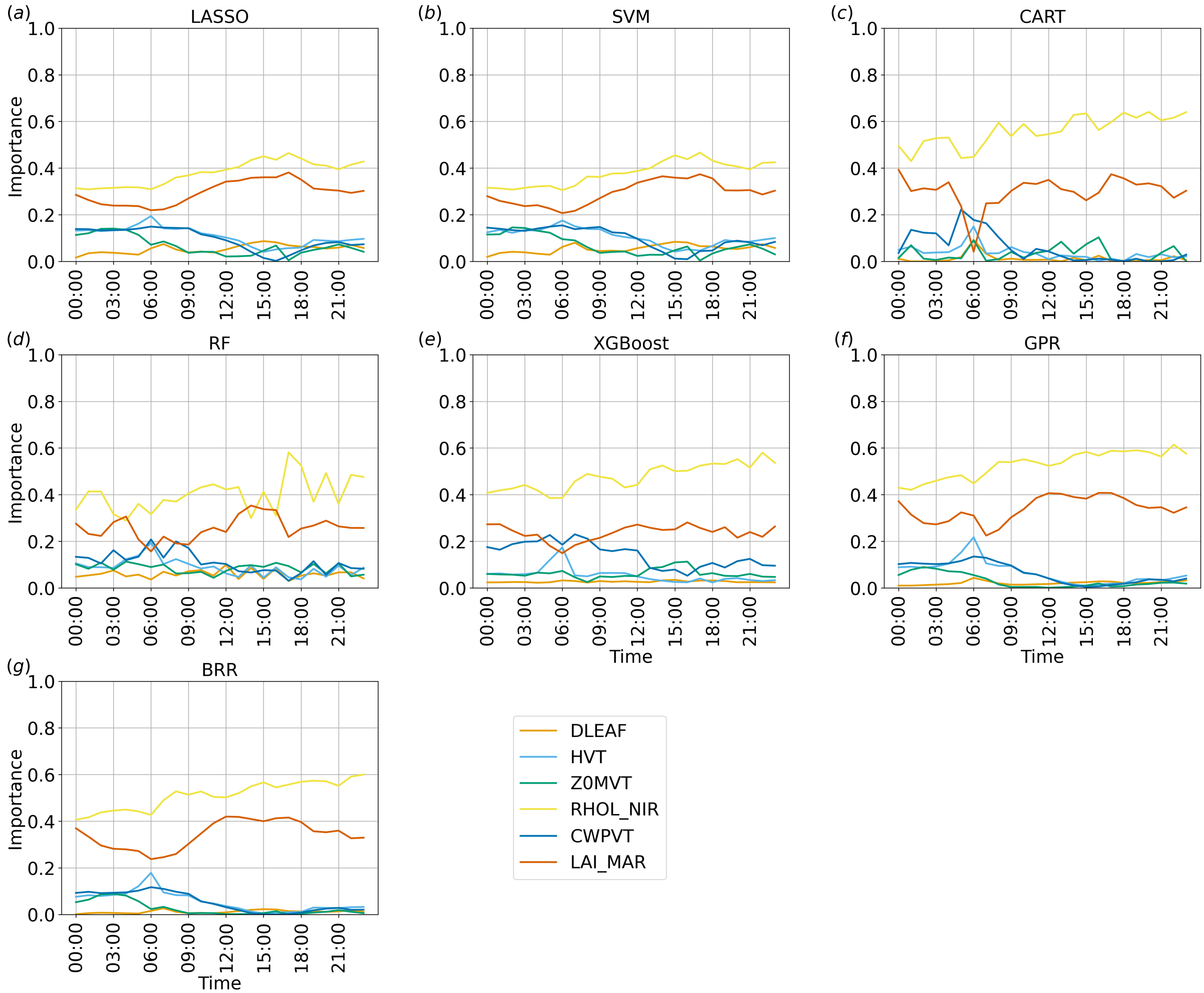

Figure 12. Time series of the importance of the parameters considered for each method implemented, considering v over water at the first vertical level.

particularly ~~relevant for this kind of circulation~~significant for this case study. Among the selected parameters, RHOL_NIR plays a central role in the main radiative processes in Noah-MP, modulating the overall canopy albedo, defining the scattered fraction of leaf intercepted radiation, and ultimately entering the computation of all radiation fluxes. LAI is involved in important processes, such as determining the canopy gaps, the fraction of vegetation exposed to sunlight, and significantly affects both sensible and latent heat fluxes, as well as the leaf boundary resistance. Although HVT might be expected to be more important due to its influence on radiation and heat trapping, its importance is probably limited by the low canopy height in the selected grassland vegetation class. CWPVT, which enters the canopy wind extinction computation, and DLEAF, which mainly affects leaf boundary resistance, were expected to play a minor role in this setup with respect to the other parameters, mainly due to their secondary role in Noah-MP.

It is interesting to note that the decision-tree-based algorithms, CART, RF and XGboost, overall detect minor differences between the less relevant parameters, while the other methods, GPR, BRR, LASSO and SVM, enhance the differences and define a clearer ranking. The reason for these differences is reasonably due to the fact that, as mentioned in Section 2, the decision-tree-based algorithms are less strict about feature shrinkage compared to other methods containing a regularization term like LASSO, hence resulting in a less clear ranking in feature importance with respect to the other methods. However, the relative importance between parameters is overall conserved, i.e. the feature importance ranking is mostly the same as the other methods for the entire length of the simulation.

It is also worth noting that, considering the importance time series obtained from GPR and BRR surrogate models, the surrogate Sobol first index agrees very well with the feature importance scores of the other algorithms, which indicates that the Sobol indices derived from BRR and GPR and the ~~regression coefficients~~ feature importance derived from the other methods have equivalent sensitivity estimation capability when convergence is properly achieved.

4.5 Vertical variability

Figure 13 and 14 show, for the land and water regions respectively, the variations in the feature importance in the lowest 10 vertical levels at different times. Since, as highlighted in the previous section, GPR is one of the methods presenting the best performance metrics, it has been chosen as the reference algorithm for this analysis. However, the results obtained with GPR are consistent with those obtained with all the other methods, particularly with LASSO, SVM and BRR.

Over the water region, the parameters' ranking does not show significant variations with height, except in correspondence with the nocturnal wind peak (06:00 UTC), when LAI_MAR becomes more important than ~~RHO-NIR~~ RHOL_NIR above the 8th model level. The situation is more complex over the land region, with more significant variations of the parameter importance with height. In particular, it can be seen that Z0MVT is more important close to the surface, especially when wind speed is stronger (06:00 UTC and 18:00 UTC), consistent with the results shown in Fig. 11, and remarking that friction ~~is particularly important~~ affects the results especially close to the surface. ~~The~~ Conversely to the decreasing vertical importance of Z0MVT, the importance of LAI_MAR and RHOL_NIR tends to increase with height ~~, especially during the night, while the importance of RHOL-NIR slightly decreases with height with the northerly land breeze and strongly increases in the lowest vertical levels when the southerly sea breeze is well developed~~(Figure 13). The vertical importance ranking converges to the water region scenario shown in Figure 14 above the lowest two vertical levels at 06:00 UTC and above the lowest 5-6 vertical levels at 18:00 UTC, i.e., above the height at which friction is playing the most important role. On the other hand, when the wind speed is weak, i.e., at 00:00 UTC and 12:00 UTC, the vertical profile of the parameters' importance values is similar over land and water at all the vertical levels investigated.

It is worth noting that the MSE for GPR, LASSO, BRR and SVM does not show significant variations in the lowest 10 vertical levels both over land and over water (Figure 15 and 16), meaning that the observed variations in feature importance are related to changes in the input-output relation rather than to uncertainty issues. This is also supported by the fact that the

metrics of these algorithms in Figure 9 show no deterioration associated with the pattern changes in 11, and that these patterns are consistent across all the surrogate models. A slightly higher variability is shown by RF and XGboost, while CART is the only method presenting a strong height dependence, in particular considering higher MSE values close to the surface at night over land and in correspondence with the northerly land breeze peak over water. These observations strengthen the evidence
595 that this method is not performing well in this case study.

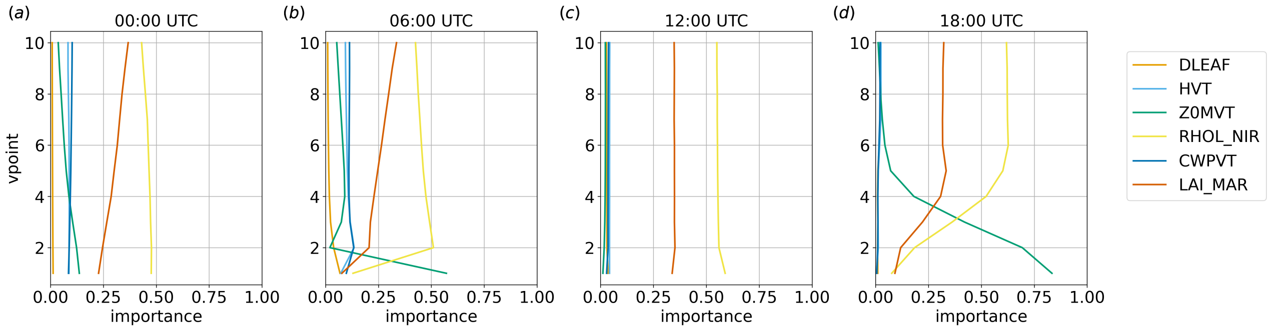

Figure 13. Parameter importance, considering v in the first 10 vertical levels over the land region at different times, for the GPR method.

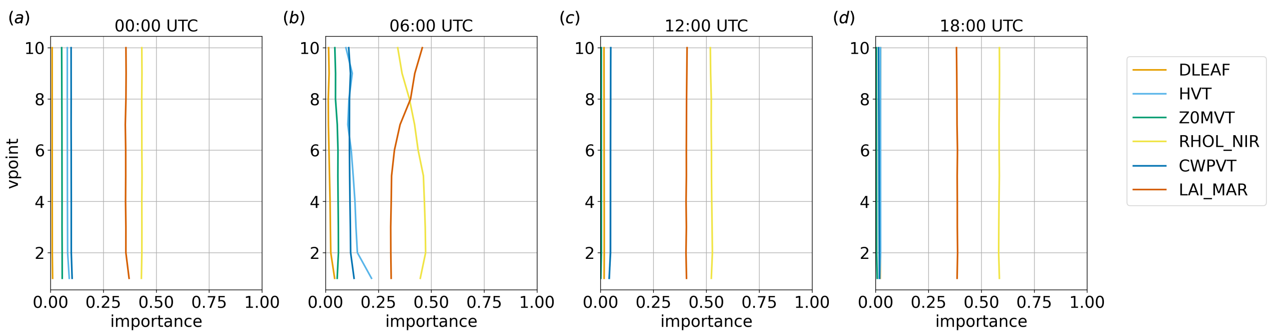

Figure 14. Parameter importance, considering v in the first 10 vertical levels over the water region at different times, for the GPR method.

5 Discussions and conclusions

This paper presented a novel automated model parameter sensitivity and importance analysis tool (ML-AMPSIT) that applies different machine learning algorithms, namely LASSO, Support Vector Machine, Classification and Decision Trees, Random Forest, Extreme Gradient Boosting, Gaussian Process Regression and Bayesian Ridge Regression, to perform sensitivity anal-
600 ysis and extract feature importance from input-output relationships. This tool was conceived to alleviate the computational burden usually associated with traditional global sensitivity analysis methods, which require a large number of model realizations, proposing an alternative approach using surrogate models or emulators. In fact, global sensitivity analysis methods,

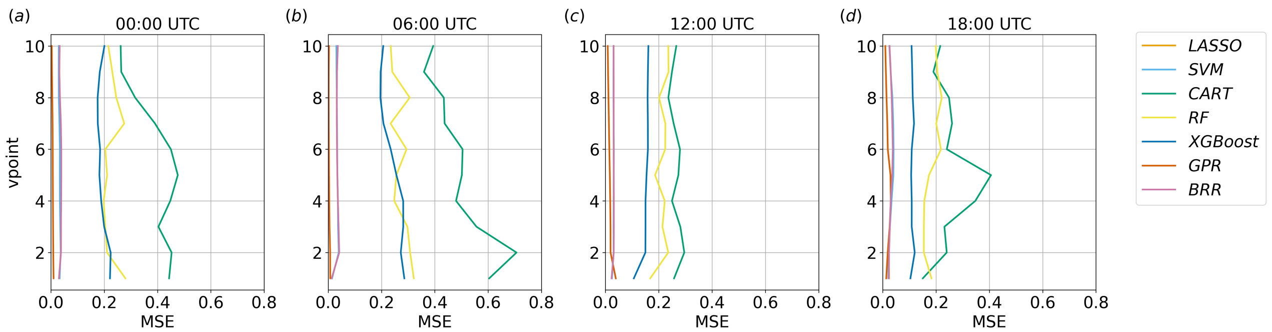

Figure 15. MSE for each method implemented, considering v in the first 10 vertical levels over the land region at different times.

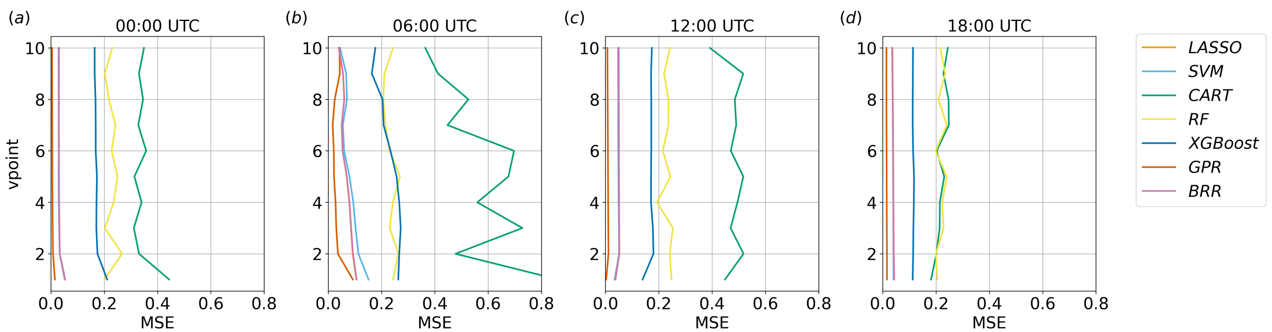

Figure 16. MSE for each method implemented, considering v in the first 10 vertical levels over the water region at different times.

such as the Sobol method, demonstrate superior performance with respect to one-at-a-time approaches, which do not consider the interaction between parameters, but the large number of model realizations needed often makes their use unfeasible for complex numerical models. On the other hand, surrogate models or emulators, trained using input-output pairs of the original high-fidelity model, offer a cost-effective means of generating accurate predictions of the output variable. The utilization of machine learning techniques provides computationally efficient solutions while considering non-linearity and interactions between variables.

The advantage of implementing different methods, also within the same family of algorithms, is multifaceted. First, if different algorithms produce consistent results, this consistency increases the reliability and robustness of the outcome. Moreover, after assessing the consistency of the results between different models of the same family, it could be more convenient to rely on the fastest method instead of the most accurate. Second, the use of different families of algorithms extends the applicability and flexibility of the tool, as their performance can vary in different scenarios.

The functionalities of the tool were tested and shown in a case study using the WRF meteorological model coupled with the Noah-MP land surface model. A sensitivity analysis applied to a set of Noah-MP parameters was presented for simulations of a sea breeze circulation over an idealized flat geometry. The different algorithms work as surrogate models of the original WRF/Noah-MP high-fidelity simulations and are able to accurately predict the original model behavior and reach robust conclusions about the parameter sensitivity given a relatively small ensemble of model runs. The efficiency of the model emulation is also tested through the computation of first-order Sobol indices from the training of Gaussian Processes Regression and Bayesian Ridge Regression, with results strongly consistent with the other proposed feature extraction methods. By integrating multiple algorithms into a flexible framework, ML-AMPSIT offers a comprehensive and reliable approach for sensitivity analysis in complex models, also allowing the evaluation of the uncertainty of the estimates by evaluating the spread between the outcomes of different algorithms.

Among different methods, Gaussian Process Regression, LASSO, Support Vector Machine, and Bayesian Ridge Regression emerged as the most reliable and robust methods. In contrast, decision-tree-based algorithms exhibited lower stability both in terms of convergence with respect to the realization number and higher uncertainty. In this case study, the linear models LASSO, Support Vector Machine, and Bayesian Ridge Regression demonstrated equal performance to the non-linearity-aware Gaussian Process Regression, suggesting the absence of strong non-linear relationships between the chosen parameters and the output variable in the analyzed domain regions.

For the best algorithms, the convergence of the feature importance was achieved with a small sample of about 20 simulations, whereas classical global sensitivity analysis approaches often require a much higher number of realizations. A qualitative comparison to evaluate the added value of ML-AMPSIT in terms of number of simulations needed to reach robust results can be performed considering two of the most advanced methods in global sensitivity analysis, i.e. the Morris method (Morris, 1991), and the Sobol method (Saltelli and Sobol', 1995), assuming to use six parameters following a Latin Hypercube Sampling (Mckay et al., 1979) with radial design (Campolongo et al., 2011). This sampling technique is one of the best trade-off for decreasing the number of simulations needed compared to a full-factorial sampling (Saltelli et al., 2008). If parameter interactions are not relevant, such as for models with low complexity and low dimensionality, a viable strategy is to use the Morris method to find out the most and least relevant parameters. For Nc centroids produced with a Latin Hypercube sampling, and k perturbations produced by a radial design, the total number of sample points, i.e. the total number of model runs required, is $N = Nc(k+1)$. A sufficient number of centroids Nc for this sampling design ranges from 10 to 50 (Campolongo et al., 2007), leading to 70-350 total simulations. However, even with this number of simulations, convergence is not guaranteed, as it depends on the specific case.

For more complex models, the Morris method can be very inefficient in stating the true parameter relevance (it is usually considered only a proxy of the true sensitivities, depending on the number of interactions and non-linearities in the model, Cuntz et al., 2015). The Sobol method is able to weigh more accurately the interaction effects between each parameter, but it is more demanding. Following Saltelli et al. (2010), to circumvent some constraints over the number of model runs required, the final number would be $N_T = N_c(k+1)(k+2)$ which, using the previous assumptions for Nc, gives a minimum number of 1120-1680 runs. Aside from the minimum amount computed above, real applications of the Sobol methods can easily exceed this value to achieve robust results (Cuntz et al., 2015, 2016). However, this is not usually feasible for complex and computationally-intensive models such as the WRF model.

It is then clear that ML-AMPSIT significantly reduces the number of simulations needed for sensitivity analysis and extraction of feature importance. Considering that all the proposed regression methods in ML-AMPSIT intrinsically account for interactions between parameters, this highlights its added value over classical global sensitivity analysis methods and points out its possible applications, especially in cases when the use of classical global sensitivity analysis methods is not feasible. Furthermore, the intercomparison of the results from different algorithms in ML-AMPSIT can reveal useful physical insights into model simulations.

It should be noted that the results presented in this paper are limited to the simple case study considered here to test the tool functionalities. In particular, it is expected that more simulations can be needed for training the algorithms in more complex scenarios, when non-linearities are more strongly involved in the input-output relations.

Code availability. The code of the ML-AMPSIT tool, along with detailed instructions on how to use it, is available at http://dx.doi.org/10.5281/zenodo.10789930 (Di Santo, 2024). The data extracted from the simulations and used for training the machine learning algorithms and producing the results presented in this paper are available at https://dx.doi.org/10.5281/zenodo.11184569. This work used WRF version 4.4 (doi:10.5065/D6MK6B4K), which includes a built-in version of Noah-MP 4.4.

Appendix A: ML-AMPSIT configuration file


```
{

  "comment1": "following lines are for generating realizations and populating the MPTABLES. folder is the reference simulation folder",
  "folder": "foldersim",
  "vegtype": 10,
  "totalsim": 100,
  "parameter_names": [ "DLEAF", "HVT", "Z0MVT", "RHOL_NIR", "CWPVT", "LAI_MAR"],

  "comment2": "for each parameters, the following matrix must contain: [middle point, percentage of perturbation]",
  "MATRIX": [
    [0.040, 50.000],
    [1.117, 50.000],
    [0.112, 50.000],
    [0.369, 50.000],
    [1.923, 50.000],
    [0.237, 50.000]
  ],

  "comment3": "following lines are specifications for ML-AMPSIT and post-processing",
  "input_pathname": "path/to/simulations/",
  "output_pathname": "path/to/output/",
  "totalhours": 36,
  "variables": ["V_MEAN","TH_MEAN"],
  "is_3d": [1,1],
  "regions": ["land", "water"],
  "verticalmax": 10,
  "tun_iter" : 10,

  "comment4": "following lines are for WRF post-processing: starting date, domain extension and points coordinate",
  "startTime": "2015-03-20 13:00:00",
  "ncfile_format":"wrfout_d01_2015-03-20_13_00_00",
  "ymax": 50,
  "xmax": 50,
  "y1": 30,
  "x1": 25,
  "y2": 20,
  "x2": 25

}
```


Figure A1. An example of the configuration file for the WRF/Noah-MP model case study.

Author contributions. Contributions to this work are delineated as per the Contributor Roles Taxonomy (CRediT) as follows: conceptualization was jointly performed by DDS, LG, CH and FC. The methodology was developed by DDS, LG and CH. DDS executed the formal analysis, developed the software, and produced the visual representations. The original draft was written by DDS, editing and reviewing was performed by LG and CH. DDS and LG conducted the investigation process. LG acquired the financial support and was responsible for the project management and coordination. Supervision was carried out by LG, FC and CH. Resources were provided by FC and CH.

Competing interests. The contact author has declared that none of the authors has any competing interests

Acknowledgements. This research was funded by Euregio Science Fund (3rd Call, IPN101) of the Europaregion Euregio.

References

Alqahtani, A., Shah, M. I., Aldrees, A., and Javed, M. F.: Comparative Assessment of Individual and Ensemble Machine Learning Models for Efficient Analysis of River Water Quality, Sustainability, 14, https://doi.org/10.3390/su14031183, 2022.

Antoniadis, A., Lambert-Lacroix, S., and Poggi, J.-M.: Random forests for global sensitivity analysis: A selective review, Reliability Engineering System Safety, 206, 107 312, https://doi.org/https://doi.org/10.1016/j.ress.2020.107312, 2021.

Antonogeorgos, G., Panagiotakos, D. B., Priftis, K. N., and Tzonou, A.: Logistic Regression and Linear Discriminant Analyses in Evaluating Factors Associated with Asthma Prevalence among 10- to 12-Years-Old Children: Divergence and Similarity of the Two Statistical Methods, International Journal of Pediatrics, 2009, 952 042, https://doi.org/10.1155/2009/952042, 2009.

Arpaci, A., Malowerschnig, B., Sass, O., and Vacik, H.: Using multi variate data mining techniques for estimating fire susceptibility of Tyrolean forests, Applied Geography, 53, 258–270, https://doi.org/https://doi.org/10.1016/j.apgeog.2014.05.015, 2014.

Arsenault, K. R., Nearing, G. S., Wang, S., Yatheendradas, S., and Peters-Lidard, C. D.: Parameter Sensitivity of the Noah-MP Land Surface Model with Dynamic Vegetation, Journal of Hydrometeorology, 19, 815 – 830, https://doi.org/10.1175/jhm-d-17-0205.1, 2018.

Baki, H., Chinta, S., Balaji, C., and Srinivasan, B.: Determining the sensitive parameters of the Weather Research and Forecasting (WRF) model for the simulation of tropical cyclones in the Bay of Bengal using global sensitivity analysis and machine learning, Geoscientific Model Development, 15, 2133–2155, https://doi.org/10.5194/gmd-15-2133-2022, 2022.

Bar Massada, A., Syphard, A. D., Stewart, S. I., and Radeloff, V. C.: Wildfire ignition-distribution modelling: a comparative study in the Huron–Manistee National Forest, Michigan, USA, International Journal of Wildland Fire, 22, 174–183, https://doi.org/10.1071/WF11178, 2013.

Bocquet, M., Brajard, J., Carrassi, A., and Bertino, L.: Bayesian inference of chaotic dynamics by merging data assimilation, machine learning and expectation-maximization, Foundations of Data Science, 2, 55–80, https://doi.org/10.3934/fods.2020004, 2020.

Bonavita, M. and Laloyaux, P.: Machine Learning for Model Error Inference and Correction, Journal of Advances in Modeling Earth Systems, 12, e2020MS002 232, https://doi.org/https://doi.org/10.1029/2020MS002232, e2020MS002232 10.1029/2020MS002232, 2020.

Boser, B. E., Guyon, I. M., and Vapnik, V. N.: A training algorithm for optimal margin classifiers, p. 144–152, https://doi.org/10.1145/130385.130401, 1992.

Box, G. E. and Tiao, G. C.: John Wiley Sons, Ltd, https://doi.org/https://doi.org/10.1002/9781118033197, 1992.

Bratley, P. and Fox, B. L.: Algorithm 659: Implementing Sobol's quasirandom sequence generator, ACM Trans. Math. Softw., 14, 88–100, https://doi.org/10.1145/42288.214372, 1988.

Breiman, L.: Random Forests, Machine Learning, 45, 5–32, https://api.semanticscholar.org/CorpusID:89141, 2001.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J.: Classification and Regression Trees, Biometrics, 40, 874, https://api.semanticscholar.org/CorpusID:29458883, 1984.

Brunton, S. L. and Kutz, J. N.: Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control, Cambridge University Press, https://doi.org/10.1017/9781108380690, 2019.

Campolongo, F., Cariboni, J., and Saltelli, A.: An effective screening design for sensitivity analysis of large models, Environmental Modelling Software, 22, 1509–1518, https://doi.org/https://doi.org/10.1016/j.envsoft.2006.10.004, modelling, computer-assisted simulations, and mapping of dangerous phenomena for hazard assessment, 2007.

Campolongo, F., Saltelli, A., and Cariboni, J.: From screening to quantitative sensitivity analysis. A unified approach, Computer Physics Communications, 182, 978–988, https://doi.org/https://doi.org/10.1016/j.cpc.2010.12.039, 2011.

Catani, F., Lagomarsino, D., Segoni, S., and Tofani, V.: Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues, Natural Hazards and Earth System Sciences, 13, 2815–2831, https://doi.org/10.5194/nhess-13-2815-2013, 2013.

Chase, R. J., Harrison, D. R., Burke, A., Lackmann, G. M., and McGovern, A.: A Machine Learning Tutorial for Operational Meteorology. Part I: Traditional Machine Learning, Weather and Forecasting, 37, 1509 – 1529, https://doi.org/https://doi.org/10.1175/WAF-D-22-0070.1, 2022.

Chase, R. J., Harrison, D. R., Lackmann, G. M., and McGovern, A.: A Machine Learning Tutorial for Operational Meteorology, Part II: Neural Networks and Deep Learning, Weather and Forecasting, https://doi.org/https://doi.org/10.1175/WAF-D-22-0187.1, 2023.

Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, ACM, https://doi.org/10.1145/2939672.2939785, 2016.

Chen, T., Zhu, L., Niu, R.-q., Trinder, C. J., Peng, L., and Lei, T.: Mapping landslide susceptibility at the Three Gorges Reservoir, China, using gradient boosting decision tree, random forest and information value models, Journal of Mountain Science, 17, 670–685, https://doi.org/10.1007/s11629-019-5839-3, 2020.

Cinquegrana, D., Zollo, A. L., Montesarchio, M., and Bucchignani, E.: A Metamodel-Based Optimization of Physical Parameters of High Resolution NWP ICON-LAM over Southern Italy, Atmosphere, 14, https://doi.org/10.3390/atmos14050788, 2023.

Clark, M. P., Kavetski, D., and Fenicia, F.: Pursuing the method of multiple working hypotheses for hydrological modeling, Water Resources Research, 47, https://doi.org/https://doi.org/10.1029/2010WR009827, 2011.

Cortes, C. and Vapnik, V. N.: Support-Vector Networks, Machine Learning, 20, 273–297, https://api.semanticscholar.org/CorpusID:52874011, 1995.

Couvreux, F., Hourdin, F., Williamson, D., Roehrig, R., Volodina, V., Villefranque, N., Rio, C., Audouin, O., Salter, J., Bazile, E., Brient, F., Favot, F., Honnert, R., Lefebvre, M.-P., Madeleine, J.-B., Rodier, Q., and Xu, W.: Process-Based Climate Model Development Harnessing Machine Learning: I. A Calibration Tool for Parameterization Improvement, Journal of Advances in Modeling Earth Systems, 13, e2020MS002 217, https://doi.org/https://doi.org/10.1029/2020MS002217, e2020MS002217 2020MS002217, 2021.

Cui, C. and Wang, D.: High dimensional data regression using Lasso model and neural networks with random weights, Information Sciences, 372, 505–517, https://doi.org/https://doi.org/10.1016/j.ins.2016.08.060, 2016.

Cuntz, M., Mai, J., Zink, M., Thober, S., Kumar, R., Schäfer, D., Schrön, M., Craven, J., Rakovec, O., Spieler, D., Prykhodko, V., Dalmasso, G., Musuuza, J., Langenberg, B., Attinger, S., and Samaniego, L.: Computationally inexpensive identification of noninformative model parameters by sequential screening, Water Resources Research, 51, 6417–6441, https://doi.org/https://doi.org/10.1002/2015WR016907, 2015.

Cuntz, M., Mai, J., Samaniego, L., Clark, M., Wulfmeyer, V., Branch, O., Attinger, S., and Thober, S.: The impact of standard and hard-coded parameters on the hydrologic fluxes in the Noah-MP land surface model, Journal of Geophysical Research: Atmospheres, 121, 10,676–10,700, https://doi.org/https://doi.org/10.1002/2016JD025097, 2016.

Dagon, K., Sanderson, B., Rosie, F., and Lawrence, D.: A machine learning approach to emulation and biophysical parameter estimation with the Community Land Model, version 5, Advances in Statistical Climatology Meteorology and Oceanography, 6, 223–244, https://doi.org/10.5194/ascmo-6-223-2020, 2020.

Daviran, M., Shamekhi, M., Ghezelbash, R., and Maghsoudi, A.: Landslide susceptibility prediction using artificial neural networks, SVMs and random forest: hyperparameters tuning by genetic optimization algorithm, International Journal of Environmental Science and Technology, 20, 259–276, https://doi.org/10.1007/s13762-022-04491-3, 2023.

Dey, A., Sahoo, D. P., Kumar, R., and Remesan, R.: A multimodel ensemble machine learning approach for CMIP6 climate model projections in an Indian River basin, International Journal of Climatology, 42, 9215–9236, https://doi.org/https://doi.org/10.1002/joc.7813, 2022.

Di Santo, D.: ML-AMPSIT, https://doi.org/10.5281/zenodo.10789930, 2024.

Dudhia, J.: Numerical Study of Convection Observed during the Winter Monsoon Experiment Using a Mesoscale Two-Dimensional Model, Journal of Atmospheric Sciences, 46, 3077 – 3107, https://doi.org/10.1175/1520-0469(1989)046<3077:NSOCOD>2.0.CO;2, 1989.

Ek, M. B., Mitchell, K. E., Lin, Y., Rogers, E., Grunmann, P., Koren, V., Gayno, G., and Tarpley, J. D.: Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model, Journal of Geophysical Research: Atmospheres, 108, https://doi.org/https://doi.org/10.1029/2002JD003296, 2003.

Elia, L., Castellaro, S., Dahal, A., and Lombardo, L.: Assessing multi-hazard susceptibility to cryospheric hazards: Lesson learnt from an Alaskan example, Science of The Total Environment, 898, 165 289, https://doi.org/10.1016/j.scitotenv.2023.165289, 2023.

Engelbrecht, A. P., Cloete, I., and Zurada, J. M.: Determining the significance of input parameters using sensitivity analysis, https://doi.org/10.1007/3-540-59497-3_199, 1995.

Farooq, F., Nasir Amin, M., Khan, K., Rehan Sadiq, M., Faisal Javed, M., Aslam, F., and Alyousef, R.: A Comparative Study of Random Forest and Genetic Engineering Programming for the Prediction of Compressive Strength of High Strength Concrete (HSC), Applied Sciences, 10, https://doi.org/10.3390/app10207330, 2020.

Fernández-Godino, M. G., Park, C., Kim, N., and Haftka, R.: Review of multi-fidelity models, Advances in Computational Science and Engineering, 1, 351–400, https://doi.org/10.3934/acse.2023015, 2017.

Fischer, M., Knippertz, P., van der Linden, R., Lemburg, A., Pante, G., Proppe, C., and Marsham, J. H.: Quantifying uncertainty in simulations of the West African monsoon with the use of surrogate models, Weather and Climate Dynamics, 5, 511–536, https://doi.org/10.5194/wcd-5-511-2024, 2024.

Forrester, A., Sobester, A., and Keane, A.: Engineering Design Via Surrogate Modelling: A Practical Guide, https://doi.org/10.1002/9780470770801, 2008.

Fowler, H. J., Blenkinsop, S., and Tebaldi, C.: Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling, International Journal of Climatology, 27, 1547–1578, https://doi.org/https://doi.org/10.1002/joc.1556, 2007.

Gholampour, A., Gandomi, A. H., and Ozbakkaloglu, T.: New formulations for mechanical properties of recycled aggregate concrete using gene expression programming, Construction and Building Materials, 130, 122–145, https://doi.org/https://doi.org/10.1016/j.conbuildmat.2016.10.114, 2017.

Gigović, L., Pourghasemi, H. R., Drobnjak, S., and Bai, S.: Testing a New Ensemble Model Based on SVM and Random Forest in Forest Fire Susceptibility Assessment and Its Mapping in Serbia's Tara National Park, Forests, 10, https://doi.org/10.3390/f10050408, 2019.

Grundner, A., Beucler, T., Gentine, P., Iglesias-Suarez, F., Giorgetta, M. A., and Eyring, V.: Deep Learning Based Cloud Cover Parameterization for ICON, Journal of Advances in Modeling Earth Systems, 14, e2021MS002 959, https://doi.org/https://doi.org/10.1029/2021MS002959, e2021MS002959 2021MS002959, 2022.

Haghiabi, A. H., Nasrolahi, A. H., and Parsaie, A.: Water quality prediction using machine learning methods, Water Quality Research Journal, 53, 3–13, https://doi.org/10.2166/wqrj.2018.025, 2018.

Han, Y., Zhang, G. J., Huang, X., and Wang, Y.: A Moist Physics Parameterization Based on Deep Learning, Journal of Advances in Modeling Earth Systems, 12, e2020MS002 076, https://doi.org/https://doi.org/10.1029/2020MS002076, e2020MS002076 2020MS002076, 2020.

He, C., P. V. M. B. F. C. D. G. R. C. T. S. R. R. G.-Y. N. Z.-L. Y. D. N. and Ek, M.: The Community Noah-MP Land Surface Modeling System Technical Description Version 5.0, No. NCAR/TN-575+STR, https://doi.org/http://dx.doi.org/10.5065/ew8g-yr95, 2023.

Herman, J. D., Kollat, J. B., Reed, P. M., and Wagener, T.: Technical Note: Method of Morris effectively reduces the computational demands of global sensitivity analysis for distributed watershed models, Hydrology and Earth System Sciences, 17, 2893–2903, https://doi.org/10.5194/hess-17-2893-2013, 2013.

Hong, S.-Y., Noh, Y., and Dudhia, J.: A New Vertical Diffusion Package with an Explicit Treatment of Entrainment Processes, Monthly Weather Review, 134, 2318 – 2341, https://doi.org/10.1175/MWR3199.1, 2006.

Kalantar, B., Pradhan, B., Naghibi, S. A., Motevalli, A., and Mansor, S.: Assessment of the effects of training data selection on the landslide susceptibility mapping: a comparison between support vector machine (SVM), logistic regression (LR) and artificial neural networks (ANN), Geomatics, Natural Hazards and Risk, 9, 49–69, https://doi.org/10.1080/19475705.2017.1407368, 2018.

Kim, S. and Boukouvala, F.: Machine learning-based surrogate modeling for data-driven optimization: a comparison of subset selection for regression techniques, Optimization Letters, 14, https://doi.org/10.1007/s11590-019-01428-7, 2020.

Kok, Z. H., Mohamed Shariff, A. R., Alfatni, M. S. M., and Khairunniza-Bejo, S.: Support Vector Machine in Precision Agriculture: A review, Computers and Electronics in Agriculture, 191, 106 546, https://doi.org/https://doi.org/10.1016/j.compag.2021.106546, 2021.

Lamberti, G. and Gorlé, C.: A multi-fidelity machine learning framework to predict wind loads on buildings, Journal of Wind Engineering and Industrial Aerodynamics, 214, 104 647, https://doi.org/10.1016/j.jweia.2021.104647, 2021.

Lee, J.-H., Sameen, M. I., Pradhan, B., and Park, H.-J.: Modeling landslide susceptibility in data-scarce environments using optimized data mining and statistical methods, Geomorphology, 303, 284–298, https://doi.org/https://doi.org/10.1016/j.geomorph.2017.12.007, 2018.

Lei, T., Ng, S., and Siu, S.: Application of ANN, XGBoost, and Other ML Methods to Forecast Air Quality in Macau, Sustainability, 15, 5341, https://doi.org/10.3390/su15065341, 2023.

Leinonen, J., Nerini, D., and Berne, A.: Stochastic Super-Resolution for Downscaling Time-Evolving Atmospheric Fields With a Generative Adversarial Network, IEEE Transactions on Geoscience and Remote Sensing, 59, 7211–7223, https://doi.org/10.1109/TGRS.2020.3032790, 2021.

Li, J., Chen, F., Zhang, G., Barlage, M., Gan, Y., Xin, Y., and Wang, C.: Impacts of Land Cover and Soil Texture Uncertainty on Land Model Simulations Over the Central Tibetan Plateau, Journal of Advances in Modeling Earth Systems, 10, 2121–2146, https://doi.org/https://doi.org/10.1029/2018MS001377, 2018.

Liu, Y., Xu, P., Cao, C., Shan, B., Zhu, K., Ma, Q., Zhang, Z., and Yin, H.: A comparative evaluation of machine learning algorithms and an improved optimal model for landslide susceptibility: a case study, Geomatics, Natural Hazards and Risk, 12, 1973–2001, https://doi.org/10.1080/19475705.2021.1955018, 2021.

Longo, R., Bellemans, A., Derudi, M., and Parente, A.: A multi-fidelity framework for the estimation of the turbulent Schmidt number in the simulation of atmospheric dispersion, Building and Environment, 185, 107 066, https://doi.org/10.1016/j.buildenv.2020.107066, 2020.

Maleki, H., Sorooshian, A., Goudarzi, G., Baboli, Z., Tahmasebi Birgani, Y., and Rahmati, M.: Air pollution prediction by using an artificial neural network model, Clean Technologies and Environmental Policy, 21, 1341–1352, https://doi.org/10.1007/s10098-019-01709-w, 2019.

Maraun, D. and Widmann, M.: Statistical Downscaling and Bias Correction for Climate Research, Cambridge University Press, https://doi.org/10.1017/9781107588783, 2018.

Maroco, J., Silva, D., Rodrigues, A., Guerreiro, M., Santana, I., and de Mendonça, A.: Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests, BMC Research Notes, 4, 299, https://doi.org/10.1186/1756-0500-4-299, 2011.

Mckay, M., Beckman, R., and Conover, W.: A Comparison of Three Methods for Selecting Vales of Input Variables in the Analysis of Output From a Computer Code, Technometrics, 21, 239–245, https://doi.org/10.1080/00401706.1979.10489755, 1979.

Meenal, R., Binu, D., Ramya, K. C., Michael, P. A., Vinoth Kumar, K., Rajasekaran, E., and Sangeetha, B.: Weather Forecasting for Renewable Energy System: A Review, Archives of Computational Methods in Engineering, 29, 2875–2891, https://doi.org/10.1007/s11831-021-09695-3, 2022.

Mendoza, P. A., Clark, M. P., Barlage, M., Rajagopalan, B., Samaniego, L., Abramowitz, G., and Gupta, H.: Are we unnecessarily constraining the agility of complex process-based models?, Water Resources Research, 51, 716–728, https://doi.org/https://doi.org/10.1002/2014WR015820, 2015.

Mlawer, E. J., Taubman, S. J., Brown, P. D., Iacono, M. J., and Clough, S. A.: Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave, Journal of Geophysical Research: Atmospheres, 102, 16 663–16 682, https://doi.org/https://doi.org/10.1029/97JD00237, 1997.

Mooers, G., Pritchard, M., Beucler, T., Ott, J., Yacalis, G., Baldi, P., and Gentine, P.: Assessing the Potential of Deep Learning for Emulating Cloud Superparameterization in Climate Models With Real-Geography Boundary Conditions, Journal of Advances in Modeling Earth Systems, 13, e2020MS002 385, https://doi.org/https://doi.org/10.1029/2020MS002385, e2020MS002385 2020MS002385, 2021.

Morris, M. D.: Factorial sampling plans for preliminary computational experiments, Quality Engineering, 37, 307–310, 1991.

Murti, M. A., Junior, R., Ahmed, A. N., and Elshafie, A.: Earthquake multi-classification detection based velocity and displacement data filtering using machine learning algorithms, Scientific Reports, 12, 21 200, https://doi.org/10.1038/s41598-022-25098-1, 2022.

Muthukrishnan, R. and Rohini, R.: LASSO: A feature selection technique in predictive modeling for machine learning, in: 2016 IEEE International Conference on Advances in Computer Applications (ICACA), pp. 18–20, https://doi.org/10.1109/ICACA.2016.7887916, 2016.

Niu, G.-Y. and Yang, Z.-L.: Effects of vegetation canopy processes on snow surface energy and mass balances, Journal of Geophysical Research: Atmospheres, 109, https://doi.org/https://doi.org/10.1029/2004JD004884, 2004.

Niu, G.-Y. and Yang, Z.-L.: An observation-based formulation of snow cover fraction and its evaluation over large North American river basins, Journal of Geophysical Research: Atmospheres, 112, https://doi.org/https://doi.org/10.1029/2007JD008674, 2007.

Niu, G.-Y., Yang, Z.-L., Dickinson, R. E., Gulden, L. E., and Su, H.: Development of a simple groundwater model for use in climate models and evaluation with Gravity Recovery and Climate Experiment data, Journal of Geophysical Research: Atmospheres, 112, https://doi.org/https://doi.org/10.1029/2006JD007522, 2007.

Niu, G.-Y., Yang, Z.-L., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., Kumar, A., Manning, K., Niyogi, D., Rosero, E., Tewari, M., and Xia, Y.: The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements, Journal of Geophysical Research: Atmospheres, 116, https://doi.org/https://doi.org/10.1029/2010JD015139, 2011.

Oliveira, S., Oehler, F., San-Miguel-Ayanz, J., Camia, A., and Pereira, J. M.: Modeling spatial patterns of fire occurrence in Mediterranean Europe using Multiple Regression and Random Forest, Forest Ecology and Management, 275, 117–129, https://doi.org/https://doi.org/10.1016/j.foreco.2012.03.003, 2012.

O'Hagan, A.: Bayesian analysis of computer code outputs: A tutorial, Reliability Engineering System Safety, 91, 1290–1300, https://doi.org/https://doi.org/10.1016/j.ress.2005.11.025, the Fourth International Conference on Sensitivity Analysis of Model Output (SAMO 2004), 2006.

Palani, S., Liong, S.-Y., and Tkalich, P.: An ANN application for water quality forecasting, Marine Pollution Bulletin, 56, 1586–1597, https://doi.org/https://doi.org/10.1016/j.marpolbul.2008.05.021, 2008.

Pourtaghi, Z. S., Pourghasemi, H. R., Aretano, R., and Semeraro, T.: Investigation of general indicators influencing on forest fire and its susceptibility modeling using different data mining techniques, Ecological Indicators, 64, 72–84, https://doi.org/https://doi.org/10.1016/j.ecolind.2015.12.030, 2016.

Pradhan, B.: A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS, Computers Geosciences, 51, 350–365, https://doi.org/https://doi.org/10.1016/j.cageo.2012.08.023, 2013.

Queipo, N. V., Haftka, R. T., Shyy, W., Goel, T., Vaidyanathan, R., and Kevin Tucker, P.: Surrogate-based analysis and optimization, Progress in Aerospace Sciences, 41, 1–28, https://doi.org/https://doi.org/10.1016/j.paerosci.2005.02.001, 2005.

Rahmati, O., Pourghasemi, H. R., and Melesse, A. M.: Application of GIS-based data driven random forest and maximum entropy models for groundwater potential mapping: A case study at Mehran Region, Iran, CATENA, 137, 360–372, https://doi.org/https://doi.org/10.1016/j.catena.2015.10.010, 2016.

Rasmussen, C. E. and Williams, C. K. I.: Gaussian Processes for Machine Learning, The MIT Press, https://doi.org/10.7551/mitpress/3206.001.0001, 2005.

Rasp, S., Pritchard, M. S., and Gentine, P.: Deep learning to represent subgrid processes in climate models, Proceedings of the National Academy of Sciences, 115, 9684–9689, https://doi.org/10.1073/pnas.1810286115, 2018.

Ren, X., Li, X., Ren, K., Song, J., Xu, Z., Deng, K., and Wang, X.: Deep Learning-Based Weather Prediction: A Survey, Big Data Research, 23, 100 178, https://doi.org/10.1016/j.bdr.2020.100178, 2020.

Ren, Y., Zhang, L., and Suganthan, P.: Ensemble Classification and Regression-Recent Developments, Applications and Future Directions [Review Article], IEEE Computational Intelligence Magazine, 11, 41–53, https://doi.org/10.1109/MCI.2015.2471235, 2016.

Rodriguez-Galiano, V., Mendes, M. P., Garcia-Soldado, M. J., Chica-Olmo, M., and Ribeiro, L.: Predictive modeling of groundwater nitrate pollution using Random Forest and multisource variables related to intrinsic and specific vulnerability: A case study in an agricultural setting (Southern Spain), Science of The Total Environment, 476-477, 189–206, https://doi.org/https://doi.org/10.1016/j.scitotenv.2014.01.001, 2014.

Ross, A., Li, Z., Perezhogin, P., Fernandez-Granda, C., and Zanna, L.: Benchmarking of Machine Learning Ocean Subgrid Parameterizations in an Idealized Model, Journal of Advances in Modeling Earth Systems, 15, e2022MS003 258, https://doi.org/https://doi.org/10.1029/2022MS003258, e2022MS003258 2022MS003258, 2023.

Sakaguchi, K. and Zeng, X.: Effects of soil wetness, plant litter, and under-canopy atmospheric stability on ground evaporation in the Community Land Model (CLM3.5), Journal of Geophysical Research: Atmospheres, 114, https://doi.org/https://doi.org/10.1029/2008JD010834, 2009.

Salmasi, F., Nouri, M., Sihag, P., and Abraham, J.: Application of SVM, ANN, GRNN, RF, GP and RT models for predicting discharge coefficients of oblique sluice gates using experimental data, Water Supply, 21, 232–248, https://doi.org/10.2166/ws.2020.226, 2020.

Saltelli, A.: Elementary Effects Method, chap. 3, pp. 109–154, John Wiley Sons, Ltd, https://doi.org/https://doi.org/10.1002/9780470725184.ch3, 2007.

Saltelli, A. and Sobol', I.: Sensitivity analysis for nonlinear mathematical models: Numerical experience, Matematicheskoe Modelirovanie, 7, 1995.

Saltelli, A., Ratto, M., Terry, A., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S.: Global Sensitivity Analysis: The Primer, https://api.semanticscholar.org/CorpusID:115957810, 2008.

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., and Tarantola, S.: Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index, Computer Physics Communications, 181, 259–270, https://doi.org/https://doi.org/10.1016/j.cpc.2009.09.018, 2010.

Sarkar, A. and Pandey, P.: River Water Quality Modelling Using Artificial Neural Network Technique, Aquatic Procedia, 4, 1070–1077, https://doi.org/https://doi.org/10.1016/j.aqpro.2015.02.135, iNTERNATIONAL CONFERENCE ON WATER RESOURCES, COASTAL AND OCEAN ENGINEERING (ICWRCOE'15), 2015.

Satir, O., Berberoglu, S., and Donmez, C.: Mapping regional forest fire probability using artificial neural network model in a Mediterranean forest ecosystem, Geomatics, Natural Hazards and Risk, 7, 1645–1658, https://doi.org/10.1080/19475705.2015.1084541, 2016.

Schaake, J. C., Koren, V. I., Duan, Q.-Y., Mitchell, K., and Chen, F.: Simple water balance model for estimating runoff at different spatial and temporal scales, Journal of Geophysical Research: Atmospheres, 101, 7461–7475, https://doi.org/https://doi.org/10.1029/95JD02892, 1996.

Schneider, R., Bonavita, M., Geer, A., Arcucci, R., Dueben, P., Vitolo, C., Le Saux, B., Demir, B., and Mathieu, P.-P.: ESA-ECMWF Report on recent progress and research directions in machine learning for Earth System observation and prediction, npj Climate and Atmospheric Science, 5, 51, https://doi.org/10.1038/s41612-022-00269-z, 2022.

Schultz, M. G., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L. H., Mozaffari, A., and Stadtler, S.: Can deep learning beat numerical weather prediction?, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 379, 20200 097, https://doi.org/10.1098/rsta.2020.0097, 2021.

Schölkopf, B. and Smola, A.: Support Vector Machines and Kernel Algorithms, Encyclopedia of Biostatistics, 5328-5335 (2005), 2002.

Sekulić, A., Kilibarda, M., Heuvelink, G. B., Nikolić, M., and Bajat, B.: Random Forest Spatial Interpolation, Remote Sensing, 12, https://doi.org/10.3390/rs12101687, 2020.

Shah, M. I., Alaloul, W. S., Alqahtani, A., Aldrees, A., Musarat, M. A., and Javed, M. F.: Predictive Modeling Approach for Surface Water Quality: Development and Comparison of Machine Learning Models, Sustainability, 13, https://doi.org/10.3390/su13147515, 2021.

Sharma, A., Paliwal, K. K., Imoto, S., and Miyano, S.: A feature selection method using improved regularized linear discriminant analysis, Machine Vision and Applications, 25, 775–786, https://doi.org/10.1007/s00138-013-0577-y, 2014.

Shen, K.-Q., Ong, C.-J., Li, X.-P., and Wilder-Smith, E. P. V.: Feature selection via sensitivity analysis of SVM probabilistic outputs, Machine Learning, 70, 1–20, https://doi.org/10.1007/s10994-007-5025-7, 2008.

Sihag, P., Kumar, V., Afghan, F. R., Pandhiani, S. M., and Keshavarzi, A.: Predictive modeling of PM2.5 using soft computing techniques: case study—Faridabad, Haryana, India, Air Quality, Atmosphere & Health, 12, 1511–1520, https://doi.org/10.1007/s11869-019-00755-z, 2019.

Skamarock, W. C., Klemp, J., Dudhia, J., Gill, D., Liu, Z., Berner, J., Wanh, W., Powers, J., Duda, M., Barker, D., and Huang, X.-Y.: A Description of the Advanced Research WRF Model Version 4.3, No. NCAR/TN-556+STR, https://doi.org/10.5065/1dfh-6p97, 2021.

Sridhara, S., Manoj, K. N., Gopakkali, P., Kashyap, G. R., Das, B., Singh, K. K., and Srivastava, A. K.: Evaluation of machine learning approaches for prediction of pigeon pea yield based on weather parameters in India, International Journal of Biometeorology, 67, 165–180, https://doi.org/10.1007/s00484-022-02396-x, 2023.

Stein, M. L.: Interpolation of spatial data: some theory for kriging, Springer Science & Business Media, 1999.

Suárez Sánchez, A., García Nieto, P., Riesgo Fernández, P., del Coz Díaz, J., and Iglesias-Rodríguez, F.: Application of an SVM-based regression model to the air quality study at local scale in the Avilés urban area (Spain), Mathematical and Computer Modelling, 54, 1453–1466, https://doi.org/https://doi.org/10.1016/j.mcm.2011.04.017, 2011.

Tian, W.: A review of sensitivity analysis methods in building energy analysis, Renewable and Sustainable Energy Reviews, 20, 411–419, https://doi.org/https://doi.org/10.1016/j.rser.2012.12.014, 2013.

Tibshirani, R.: Regression Shrinkage and Selection Via the Lasso, Journal of the Royal Statistical Society: Series B (Methodological), 58, 267–288, https://doi.org/https://doi.org/10.1111/j.2517-6161.1996.tb02080.x, 1996.

Torres, M.: A Machine Learning Method for Parameter Estimation and Sensitivity Analysis, pp. 330–343, https://doi.org/10.1007/978-3-030-77977-1_26, 2021.

Trabelsi, F. and Bel Hadj Ali, S.: Exploring Machine Learning Models in Predicting Irrigation Groundwater Quality Indices for Effective Decision Making in Medjerda River Basin, Tunisia, Sustainability, 14, https://doi.org/10.3390/su14042341, 2022.

Vapnik, V.N. Chervonenkis, A.: On a class of algorithms of learning pattern recognition, 25, 112–120, 1963.

Watson-Parris, D., Williams, A., Deaconu, L., and Stier, P.: Model calibration using ESEm v1.1.0 – an open, scalable Earth system emulator, Geoscientific Model Development, 14, 7659–7672, https://doi.org/10.5194/gmd-14-7659-2021, 2021.

Wicker, L. and Skamarock, W.: Time-Splitting Methods for Elastic Models Using Forward Time Schemes, Monthly Weather Review - MON WEATHER REV, 130, https://doi.org/10.1175/1520-0493(2002)130<2088:TSMFEM>2.0.CO;2, 2002.

Williamson, D., Goldstein, M., Allison, L., Blaker, A., Challenor, P., Jackson, L., and Yamazaki, K.: History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble, Climate Dynamics, 41, 1703–1729, https://doi.org/10.1007/s00382-013-1896-4, 2013.

Wolff, B., Kramer, O., and Heinemann, D.: Selection of Numerical Weather Forecast Features for PV Power Predictions with Random Forests, in: Data Analytics for Renewable Energy Integration, edited by Woon, W. L., Aung, Z., Kramer, O., and Madnick, S., pp. 78–91, Springer International Publishing, Cham, 2017.

Wu, W., Li, M.-F., Xu, X., Tang, X.-P., Yang, C., and Liu, H.-B.: The transferability of random forest and support vector machine for estimating daily global solar radiation using sunshine duration over different climate zones, Theoretical and Applied Climatology, 146, 45–55, https://doi.org/10.1007/s00704-021-03726-6, 2021.

Wu, Z., Cui, N., Gong, D., Zhu, F., Xing, L., Zhu, B., Chen, X., Wen, S., and Liu, Q.: Simulation of daily maize evapotranspiration at different growth stages using four machine learning models in semi-humid regions of northwest China, Journal of Hydrology, 617, 128 947, https://doi.org/https://doi.org/10.1016/j.jhydrol.2022.128947, 2023.

Yang, W.-X., Wang, F.-F., Pan, Y.-Y., Xie, J.-Q., Lu, M.-H., and You, C.-G.: Comparison of ischemic stroke diagnosis models based on machine learning, Frontiers in Neurology, 13, https://doi.org/10.3389/fneur.2022.1014346, 2022.

Yang, Z.-L., Niu, G.-Y., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., Longuevergne, L., Manning, K., Niyogi, D., Tewari, M., and Xia, Y.: The community Noah land surface model with multiparameterization options (Noah-MP): 2. Evaluation over global river basins, Journal of Geophysical Research: Atmospheres, 116, https://doi.org/https://doi.org/10.1029/2010JD015140, 2011.

Yilmaz, I.: Comparison of landslide susceptibility mapping methodologies for Koyulhisar, Turkey: conditional probability, logistic regression, artificial neural networks, and support vector machine, Environmental Earth Sciences, 61, 821–836, https://doi.org/10.1007/s12665-009-0394-9, 2010.

Youssef, A. M., Pourghasemi, H. R., Pourtaghi, Z. S., and Al-Katheeri, M. M.: Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region, Saudi Arabia, Landslides, 13, 839–856, https://doi.org/10.1007/s10346-015-0614-1, 2016.

Yu, R., Yang, Y., Yang, L., Han, G., and Move, O. A.: RAQ–A Random Forest Approach for Predicting Air Quality in Urban Sensing Systems, Sensors, 16, https://doi.org/10.3390/s16010086, 2016.

Yu, Z., Dai, H., Yang, J., Zhu, Y., and Yuan, S.: Global sensitivity analysis with deep learning-based surrogate models for unraveling key parameters and processes governing redox zonation in riparian zone, Journal of Hydrology, 638, 131 442, https://doi.org/https://doi.org/10.1016/j.jhydrol.2024.131442, 2024.

Yuval, J. and O'Gorman, P. A.: Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions, Nature Communications, 11, 3295, https://doi.org/10.1038/s41467-020-17142-3, 2020.

Zhang, X., Srinivasan, R., and Van Liew, M.: Approximating SWAT Model Using Artificial Neural Network and Support Vector Machine1, JAWRA Journal of the American Water Resources Association, 45, 460–474, https://doi.org/https://doi.org/10.1111/j.1752-1688.2009.00302.x, 2009.

Zhou, C., Yin, K., Cao, Y., Ahmed, B., Li, Y., Catani, F., and Pourghasemi, H. R.: Landslide susceptibility modeling applying machine learning methods: A case study from Longju in the Three Gorges Reservoir area, China, Computers Geosciences, 112, 23–37, https://doi.org/https://doi.org/10.1016/j.cageo.2017.11.019, 2018.

Zouhri, W., Homri, L., and Dantan, J.-Y.: Handling the impact of feature uncertainties on SVM: A robust approach based on Sobol sensitivity analysis, Expert Systems with Applications, 189, 115 691, https://doi.org/https://doi.org/10.1016/j.eswa.2021.115691, 2022.