https://doi.org/10.5194/gmd-2024-56
Preprint. Discussion started: 18 April 2024

Geoscientific Model Development Discussions

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,

https://doi.org/10.5194/gmd-2024-56
Preprint. Discussion started: 18 April 2024

Ratto, 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 (W. C. 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, Zhu, 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), 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; Z. Wu et al. 2023), climatological applications (W. 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; W.-X. 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), 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, Harrison, Burke, et al. 2022; Chase, Harrison, Lackmann, et al. 2023), the extraction of feature importance remains relatively uncommon in the meteorological/climate modeling literature (X. Ren et al. 2020; Baki et al. 2022).

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, 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,

https://doi.org/10.5194/gmd-2024-56
Preprint. Discussion started: 18 April 2024


Geoscientific Model Development Discussions

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. While most of them directly provide a measure of feature importance, the last two methods are probabilistic methods, specifically used in this framework for a fast implementation of the Sobol method, leading to a computationally efficient way to obtain the Sobol sensitivity indices directly from the ML-inferred relations between input and output data.

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.

https://doi.org/10.5194/gmd-2024-56
Preprint. Discussion started: 18 April 2024

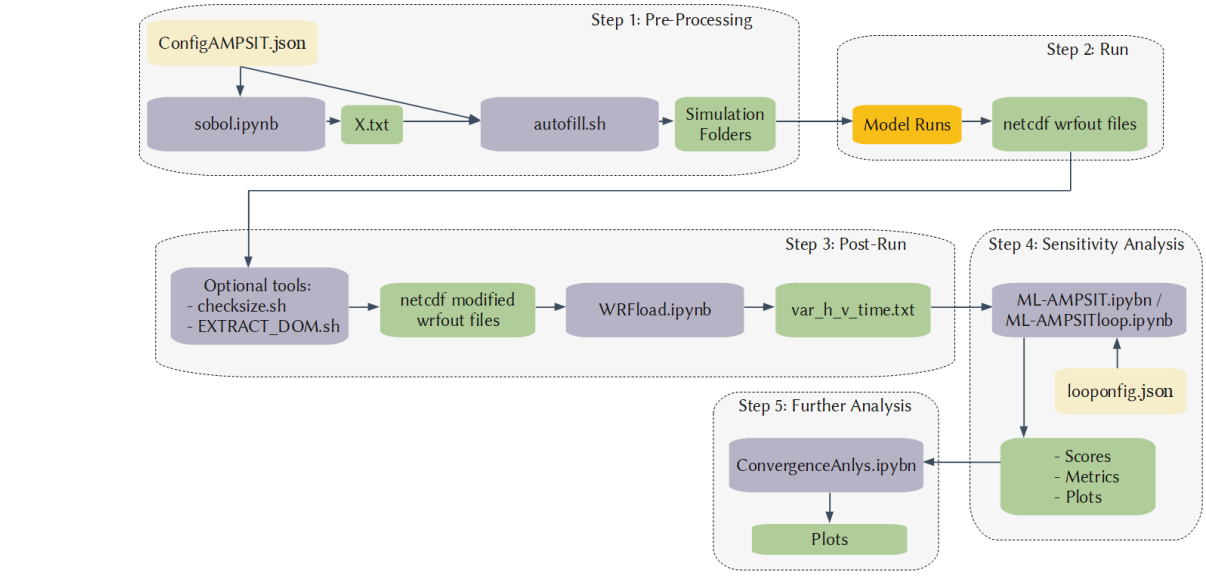


Figure 1. ML-AMPSIT workflow. The main code scripts are indicated with the blue boxes. The yellow boxes indicate the configuration files that need to be filled by the user, the green boxes refer to all the output files that eventually become inputs for other subsequent scripts, and the orange box indicates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

2 Methods

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. 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, Annoni, 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.

https://doi.org/10.5194/gmd-2024-56
Preprint. Discussion started: 18 April 2024

Geoscientific Model Development Discussions

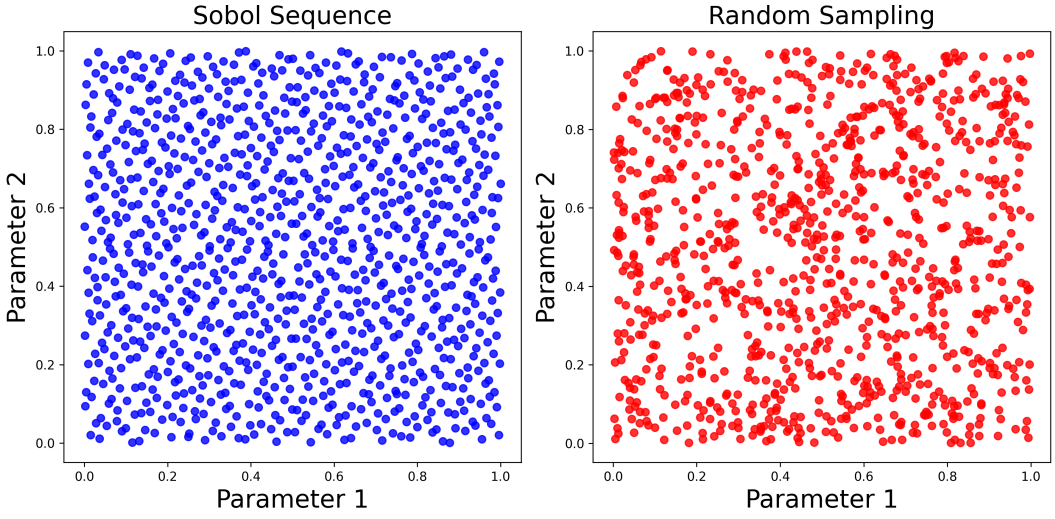


Figure 2. Demonstration of the differences between random sampling (red dots) and Sobol sampling (blue dots) in representing the parameter space. The Sobol sequence is able to more uniformly cover the parameter space, avoiding the presence of very close points, as it occurs in the random sequence.

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

https://doi.org/10.5194/gmd-2024-56
Preprint. Discussion started: 18 April 2024

Geoscientific Model Development Discussions

by the user in the configuration file. The resulting output data, which serves as input to the sensitivity analysis tool *ML-AMPSIT.ipybn*, 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.

4. The sensitivity phase is performed by the main script of the ML-AMPSIT tool *ML-AMPSIT.ipybn*, 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 (Y. 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* 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.

https://doi.org/10.5194/gmd-2024-56
Preprint. Discussion started: 18 April 2024

Geoscientific Model Development Discussions

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 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*. This script manipulates all the outputs of *ML-AMPSIT.ipybn* or *ML-AMPSITloop.ipybn* 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 ((“Elementary Effects Method” 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.

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)). With this approach, the variance $V(Y)$ is decomposed as:

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

https://doi.org/10.5194/gmd-2024-56
Preprint. Discussion started: 18 April 2024


Geoscientific Model Development Discussions

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, 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)). 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 (A. Vapnik V. C. 1963), was refined multiple times (Boser et al. 1992; Cortes and V. N. Vapnik 1995; Schölkopf and Smola 2002), and is considered a core method in ML.

https://doi.org/10.5194/gmd-2024-56
Preprint. Discussion started: 18 April 2024


Geoscientific Model Development Discussions

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.

https://doi.org/10.5194/gmd-2024-56
Preprint. Discussion started: 18 April 2024

Geoscientific Model Development Discussions

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 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 the shape and properties of the functions that the Gaussian Process can learn, 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. 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, Ratto, 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 $p(H)$ and a likelihood function $p(E|H)$, 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.

https://doi.org/10.5194/gmd-2024-56
Preprint. Discussion started: 18 April 2024


Geoscientific Model Development Discussions

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 importance coefficients is also implemented here to compute the Sobol first-order sensitivity index. Therefore, the user can compare the Sobol indices evaluated with both these two methods, providing information on their robustness and reliability.

2.3.8 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.

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 (W. C. 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 W. 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** Niu, Z.-L. Yang, et al. 2011, Z.-L. 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/).

https://doi.org/10.5194/gmd-2024-56
Preprint. Discussion started: 18 April 2024


Geoscientific Model Development Discussions

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 $\theta_s = 280$ K, $\Gamma = 3.2$ K 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 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 land-atmosphere exchange. In Noah-MP, the canopy radiative transfer scheme used is the "modified two-stream" (Niu and Z.-L. 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 with groundwater option is used for runoff and groundwater processes (Niu and Z.-L. 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.

https://doi.org/10.5194/gmd-2024-56
Preprint. Discussion started: 18 April 2024


Geoscientific Model Development Discussions

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), 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, Mai, Samaniego, 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 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. 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.

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 3 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 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.

Figure 4 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. 4a,b) is shallower than the southerly sea breeze (Fig. 4e,f). Moreover, the comparison between Fig. 4e and Fig. 4f 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 at 07:00 UTC and 19:00 UTC, while over water 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. 4c and Fig. 4d), 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

https://doi.org/10.5194/gmd-2024-56
Preprint. Discussion started: 18 April 2024

Geoscientific Model Development Discussions

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.

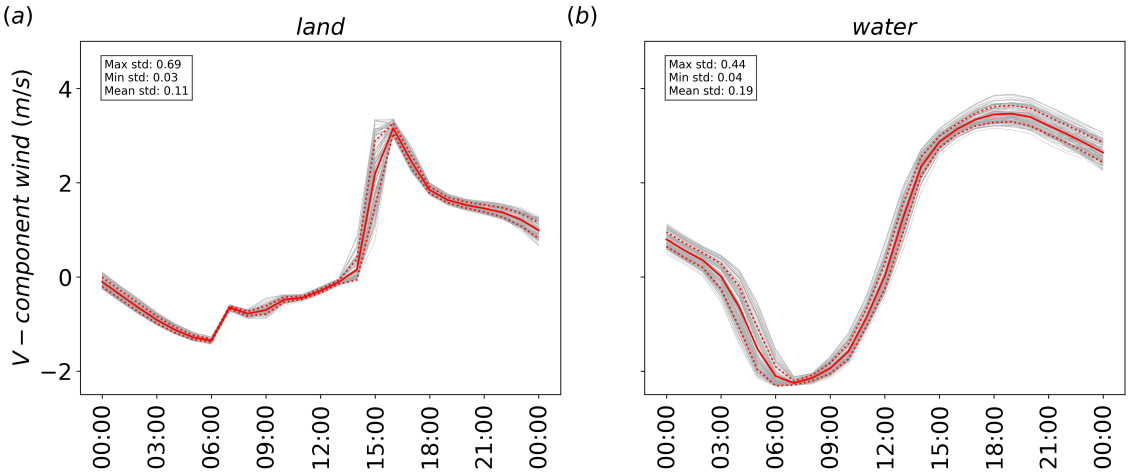


Figure 3. Time series of the south-north wind component at the first vertical level. The solid and dashed red lines represent the ensemble mean and standard deviation, respectively, while the gray lines represent the output of the single simulations. The values of the maximum and minimum hourly standard deviations (std) are also reported.

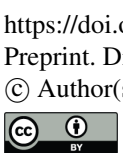
4.2 ML-AMPSIT output

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 5, 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.

4.3 Convergence analysis

Figure 6 and 7 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

https://doi.org/10.5194/gmd-2024-56
Preprint. Discussion started: 18 April 2024



Geoscientific Model Development
Discussions
Open Access
EGU

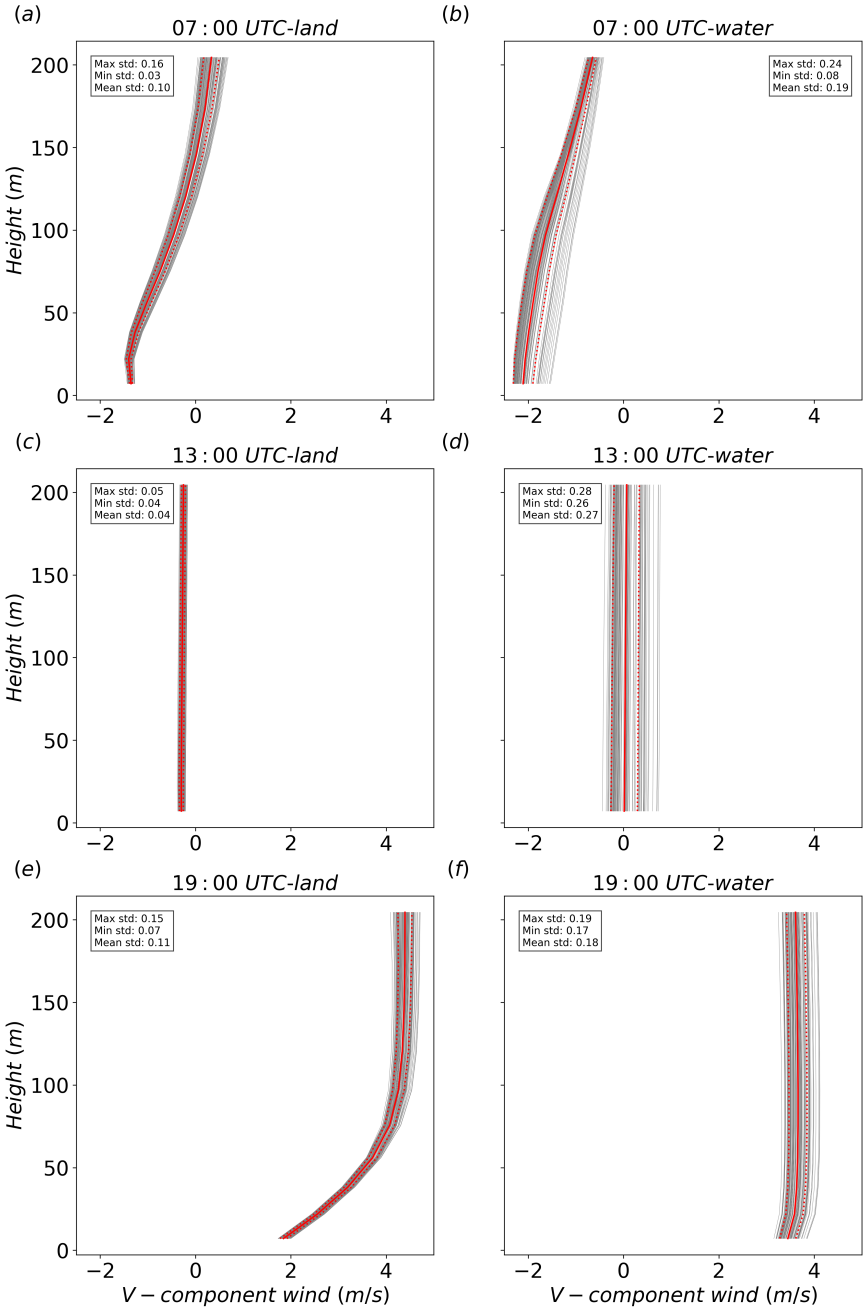


Figure 4. Vertical profiles of the south-north wind component in the lowest 200 m above the surface, at a) 07:00, b) 13:00, and c) 19:00 UTC. The solid and dashed red lines represent the ensemble mean and standard deviation, respectively, while the gray lines represent the output of the single simulations. The values of the maximum and minimum hourly standard deviations (std) are also reported.

https://doi.org/10.5194/gmd-2024-56
Preprint. Discussion started: 18 April 2024

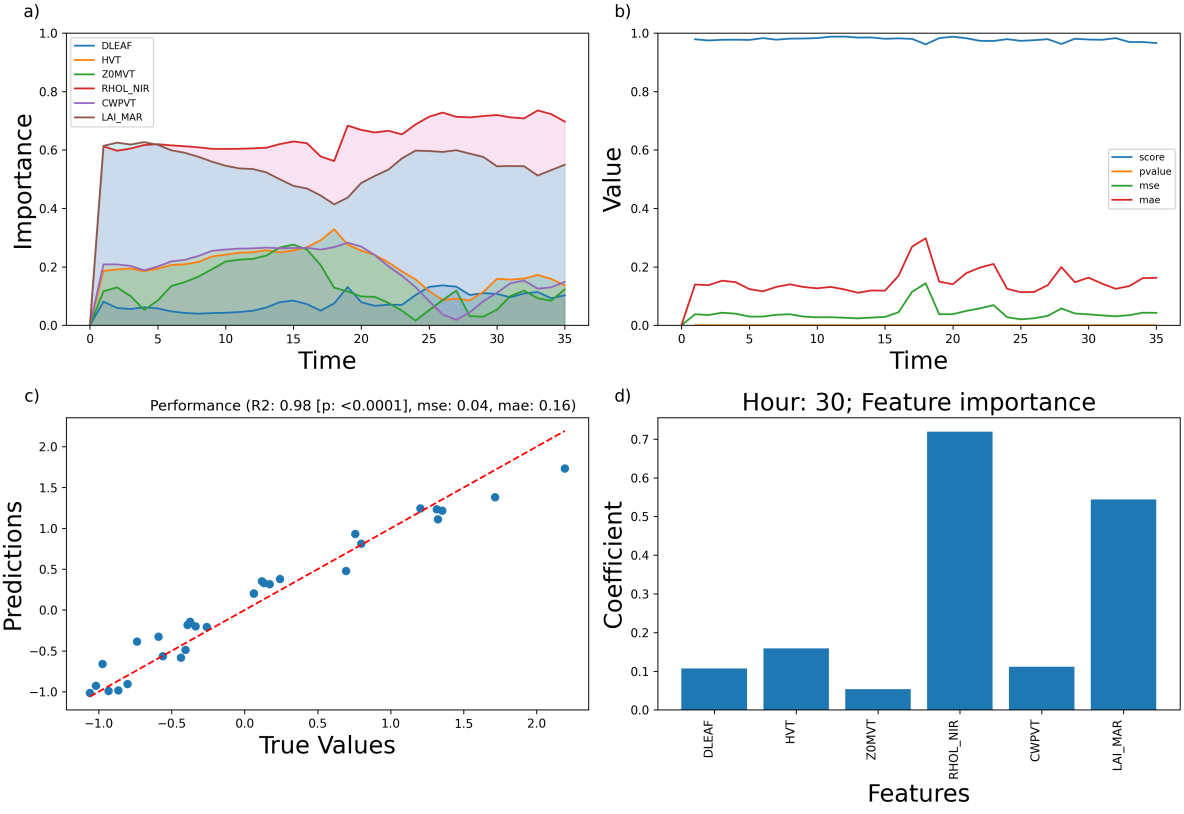


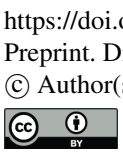
Figure 5. Example of the ML-AMPSIT output for a LASSO regression of 100 simulations with a duration of 36 hours (spin-up phase included), focusing on v at the third vertical level above the water point and displaying the metrics for the 30th simulated hour: a) importance time series of the 6 parameters; 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.

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, considering results at 13:00 UTC for the metrics convergence and at 8 different times for the feature importance. 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 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

https://doi.org/10.5194/gmd-2024-56
Preprint. Discussion started: 18 April 2024



Geoscientific Model Development Discussions

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.

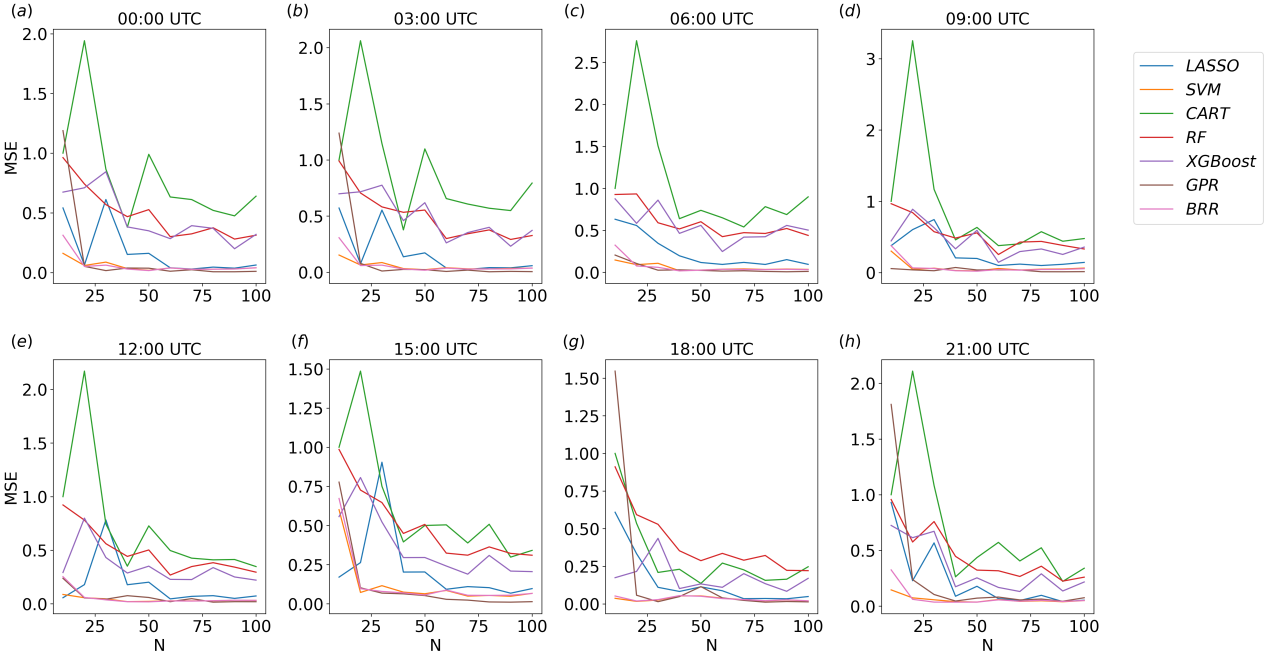


Figure 6. Convergence of the MSE with the number of realizations for each method implemented, considering v over land, at the first vertical level, at 8 different times.

4.4 Parameter importance analysis

Figures 8 and 9 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 10 and 11 show, for the same variable and regions, the time series of the feature importance for each of the proposed methods.

In this example, GPR shows the best metrics, closely followed by LASSO, SVM and BRR, 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 16:00 UTC over land and with the nocturnal peak of the northerly wind around 07:00 UTC over water. 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 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.

https://doi.org/10.5194/gmd-2024-56
Preprint. Discussion started: 18 April 2024

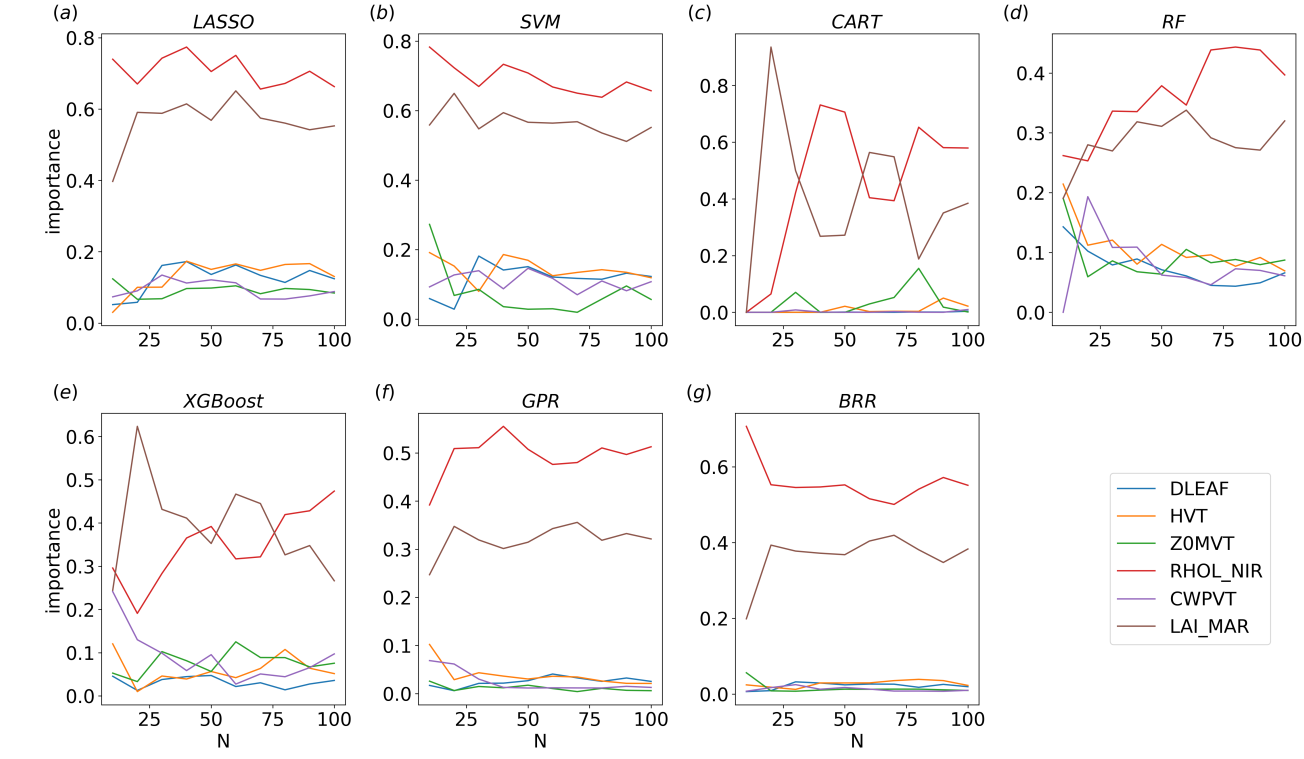


Figure 7. Convergence of the feature importance with the number of realizations, for each method implemented, considering v over land, at the first vertical level, at 13:00 UTC.

For both regions, all methods agree very well 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. 8).

Figure 10 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 alternate as the most important parameters, with RHOL-NIR dominating during the entire day, with the exception of the hours close to sunrise and sunset, when ZOMVT becomes more important. This seems to indicate a stronger role of the surface friction when stronger winds are present (ZOMVT directly influences the surface friction), coherently with the wind speed profiles shown in Figure 4a,e. LAI_MAR shows an importance almost comparable to RHOL-NIR during the day, especially with LASSO, SVM and RF, while 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 results are more uniform than over land, and the parameters' ranking does not show significant variations during the whole day. In particular, the dominant parameters are RHOL-NIR and LAI_MAR, with Z0MVT always showing

https://doi.org/10.5194/gmd-2024-56
Preprint. Discussion started: 18 April 2024

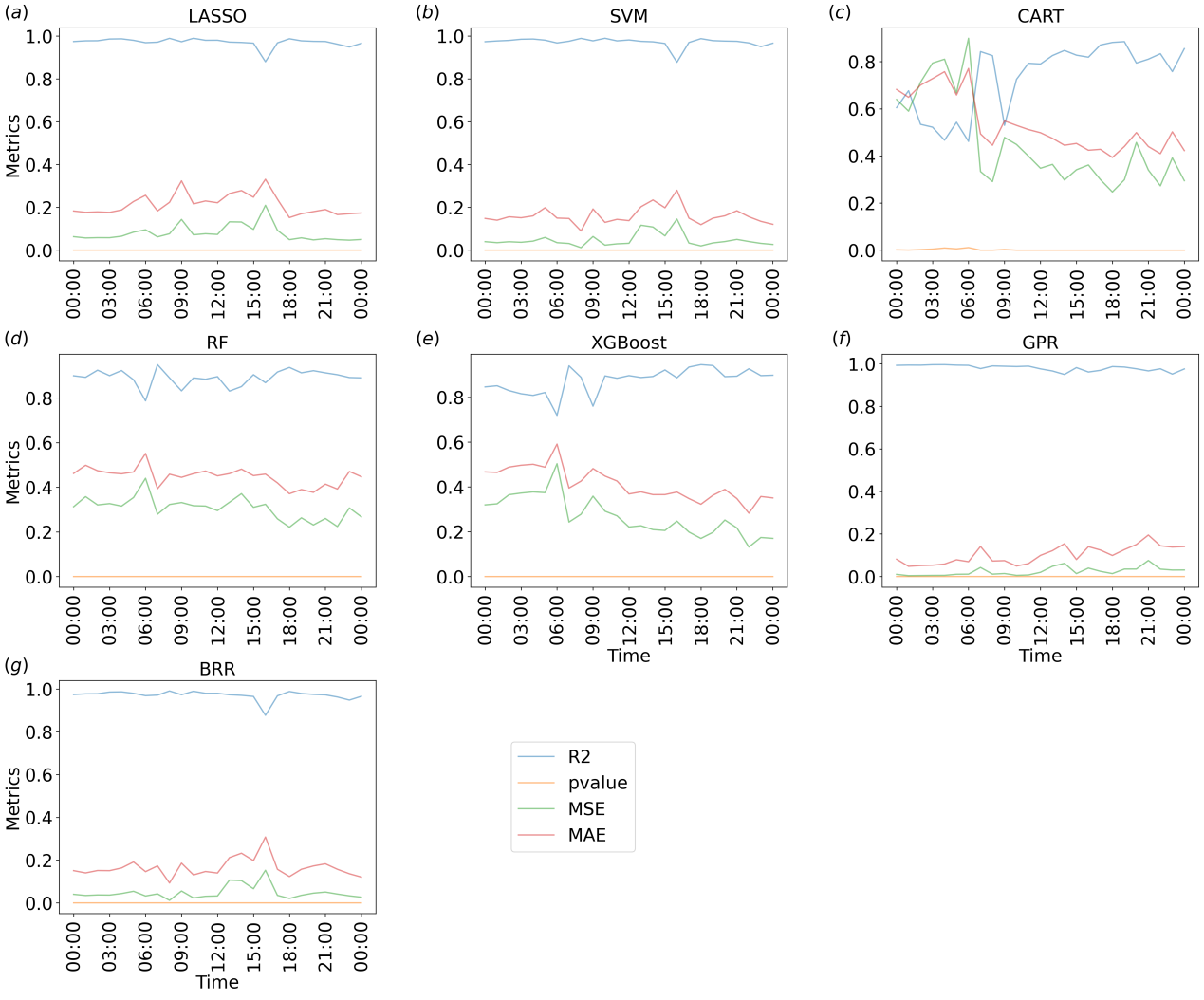


Figure 8. Time series of the performance metrics for each method implemented, considering v over land at the first vertical level: blue, green, red and orange lines represent respectively R^2, MSE, MAE and p-value.

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 particularly relevant for this kind of circulation.

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,

https://doi.org/10.5194/gmd-2024-56
Preprint. Discussion started: 18 April 2024

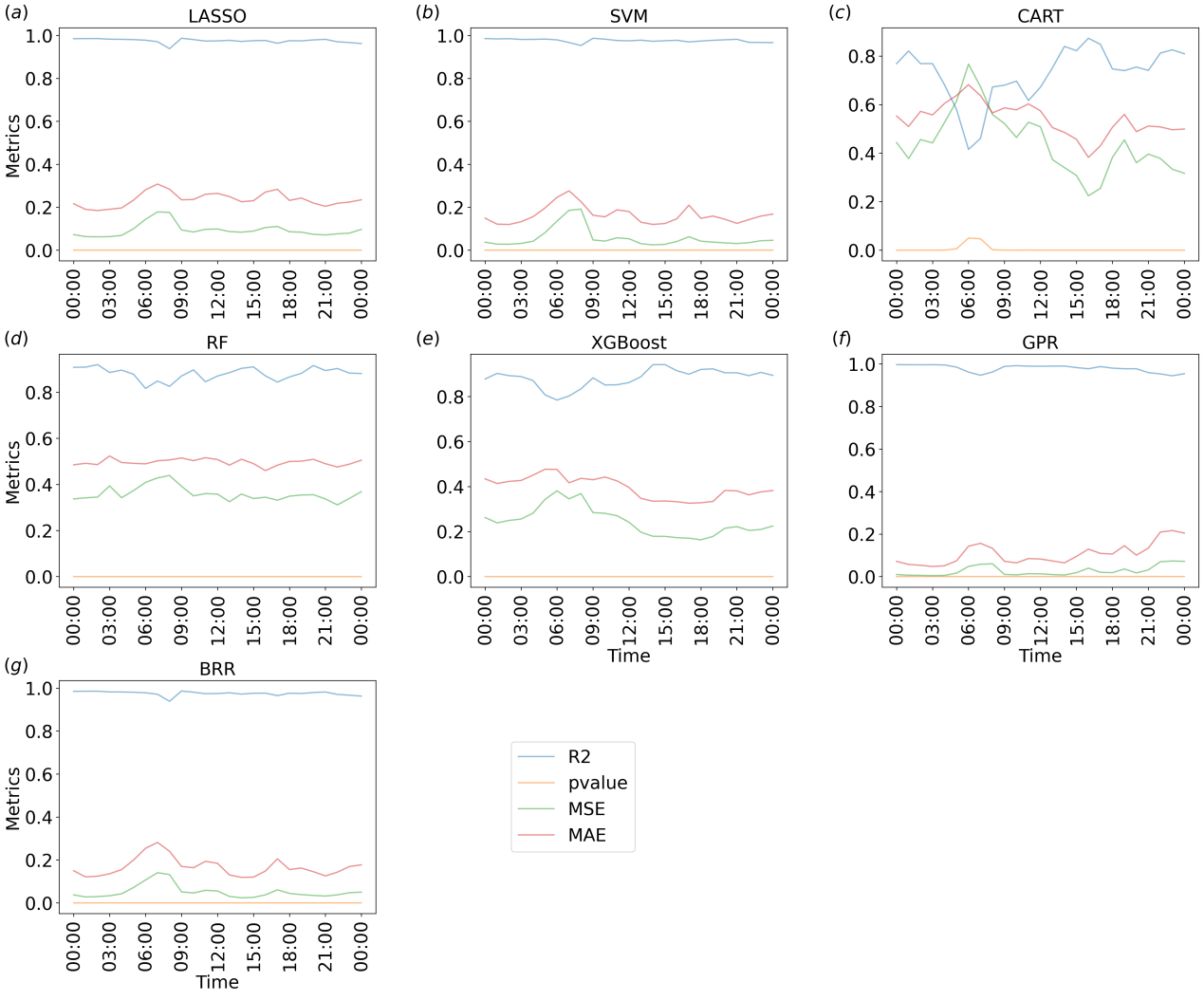


Figure 9. Time series of the performance metrics for each method implemented, considering v over water at the first vertical level: blue, green, red and orange lines represent respectivelyR^2, MSE, MAE and p-value.

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 derived from the other methods have equivalent sensitivity estimation capability when convergence is properly achieved.

https://doi.org/10.5194/gmd-2024-56
Preprint. Discussion started: 18 April 2024

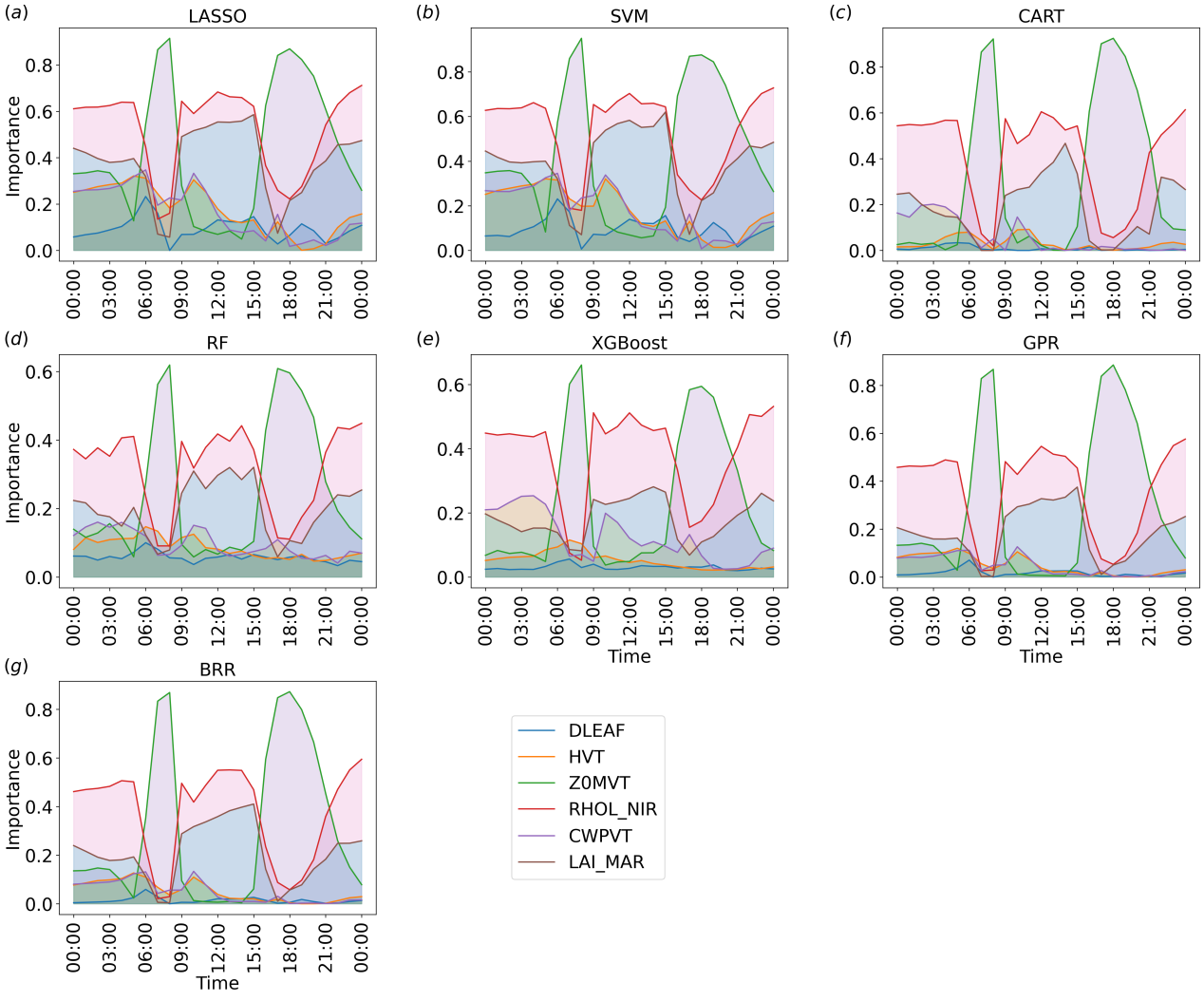


Figure 10. Time series of the importance of the parameters considered for each method implemented, considering v over land at the first vertical level.

4.5 Vertical variability

Figure 12 and 13 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 above the 8th model

https://doi.org/10.5194/gmd-2024-56
Preprint. Discussion started: 18 April 2024

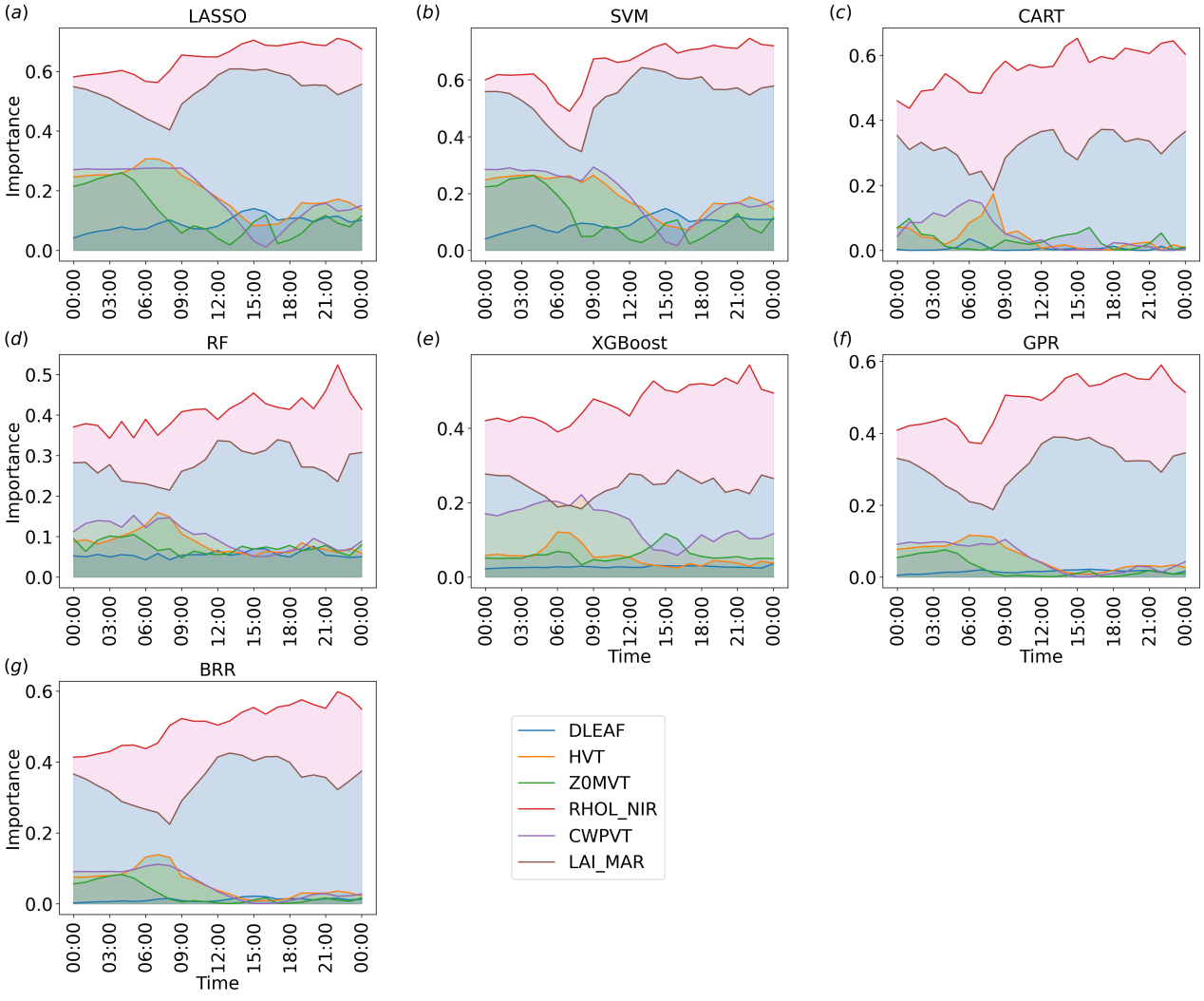
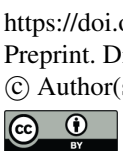


Figure 11. Time series of the importance of the parameters considered for each method implemented, considering v over water at the first vertical level.

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, consistent with the results shown in Fig. 10, and remarking that friction is particularly important close to the surface. The importance of LAI_MAR 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.

https://doi.org/10.5194/gmd-2024-56
Preprint. Discussion started: 18 April 2024



Geoscientific Model Development Discussions
Open Access
EGU

460 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 14 and 15), meaning that the observed variations in feature importance are related to changes in the input-output relation rather than to uncertainty issues. 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
465 observations strengthen the evidence that this method is not performing well in this case study.

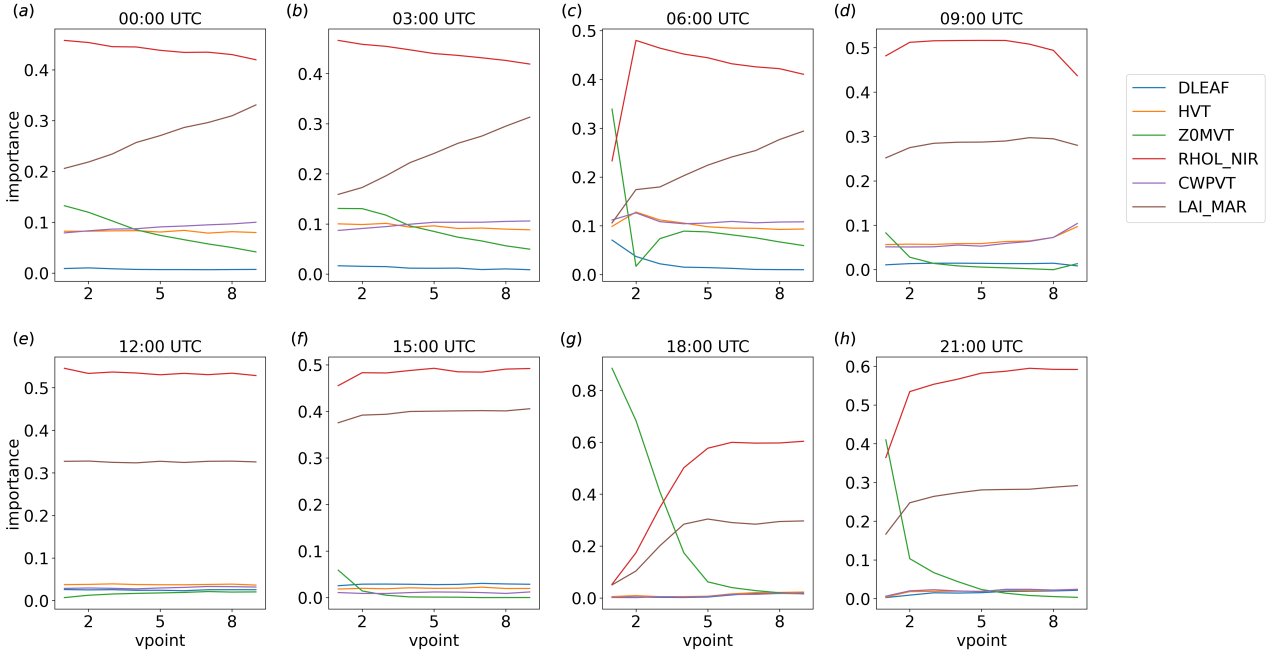


Figure 12. Parameter importance, considering v in the first 10 vertical levels over the land region at different times, for the GPR method.

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-
470 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, 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

https://doi.org/10.5194/gmd-2024-56
Preprint. Discussion started: 18 April 2024

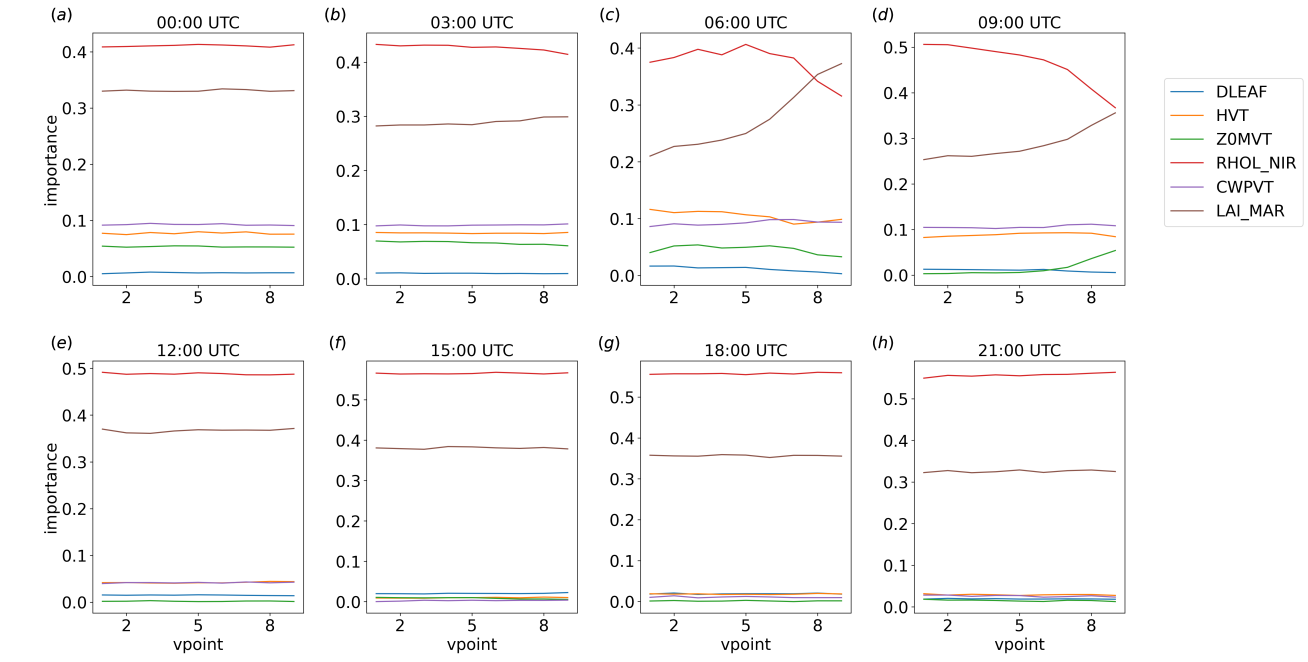


Figure 13. Parameter importance, considering v in the first 10 vertical levels over the water region at different times, for the GPR method.

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 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

https://doi.org/10.5194/gmd-2024-56
Preprint. Discussion started: 18 April 2024

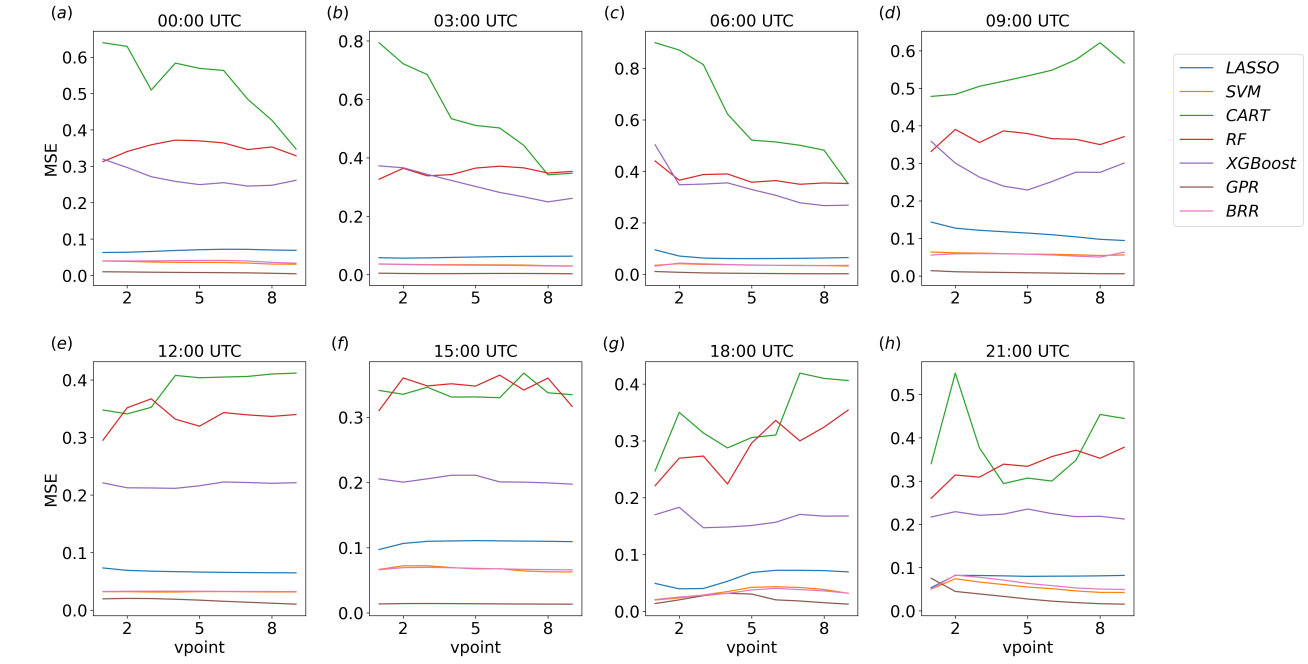


Figure 14. MSE for each method implemented, considering v in the first 10 vertical levels over the land region at different times.

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, Saltelli, 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, Ratto, 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, Cariboni, 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.

https://doi.org/10.5194/gmd-2024-56
Preprint. Discussion started: 18 April 2024

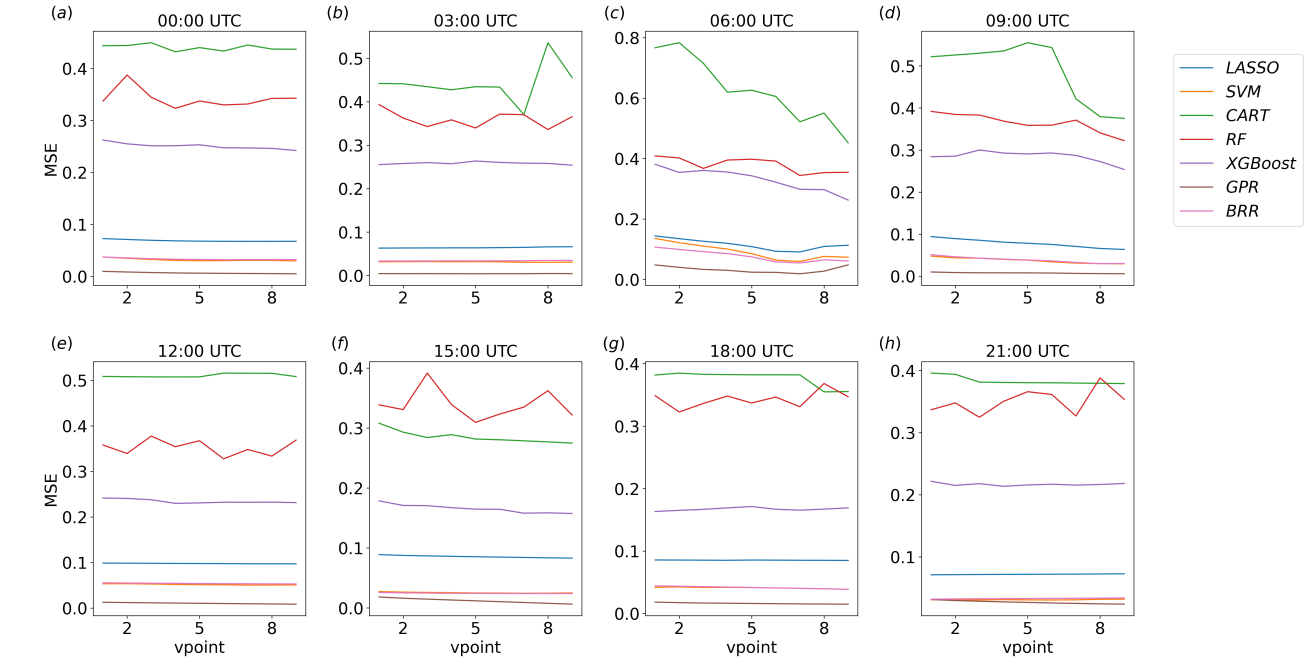


Figure 15. MSE for each method implemented, considering v in the first 10 vertical levels over the water region at different times.

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, Mai, Zink, 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, Annoni, 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, Mai, Zink, et al. 2015; Cuntz, Mai, Samaniego, et al. 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.

https://doi.org/10.5194/gmd-2024-56
Preprint. Discussion started: 18 April 2024

Geoscientific Model Development Discussions
Open Access
EGU

Code availability. The code for the ML-AMPSIT tool, along with detailed instructions on how to use it, is available at https://github.com/ML-AMPSIT/ML-AMPSIT (Di Santo 2024)

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 SAML and post-processing",
  "input_pathname": "E:/X231023FLATseabreeze/",
  "output_pathname": "C:/Users/dario/Documents/MATLABdott/WRF_IDEAL_SIMUL/VARStxt/XFLAT231023p5mean0/",
  "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 18:00:00",
  "ncfile_format":"wrfout_d01_2015-03-20_18_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

https://doi.org/10.5194/gmd-2024-56
Preprint. Discussion started: 18 April 2024

Geoscientific Model Development Discussions

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

References

Alqahtani, A. et al. (2022). "Comparative Assessment of Individual and Ensemble Machine Learning Models for Efficient Analysis of River Water Quality". In: *Sustainability* 14.3. ISSN: 2071-1050. DOI: 10.3390/su14031183. URL: https://www.mdpi.com/2071-1050/14/3/1183.

Antoniadis, A., S. Lambert-Lacroix, and J.-M. Poggi (2021). "Random forests for global sensitivity analysis: A selective review". In: *Reliability Engineering System Safety* 206, p. 107312. ISSN: 0951-8320. DOI: https://doi.org/10.1016/j.ress.2020.107312. URL: https://www.sciencedirect.com/science/article/pii/S0951832020308073.

Antonogeorgos, G. et al. (Mar. 2009). "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". In: *International Journal of Pediatrics* 2009, p. 952042. ISSN: 1687-9740. DOI: 10.1155/2009/952042. URL: https://doi.org/10.1155/2009/952042.

Arpaci, A. et al. (2014). "Using multi variate data mining techniques for estimating fire susceptibility of Tyrolean forests". In: *Applied Geography* 53, pp. 258–270. ISSN: 0143-6228. DOI: https://doi.org/10.1016/j.apgeog.2014.05.015. URL: https://www.sciencedirect.com/science/article/pii/S0143622814001106.

Arsenault, K. R. et al. (2018). "Parameter Sensitivity of the Noah-MP Land Surface Model with Dynamic Vegetation". In: *Journal of Hydrometeorology* 19.5, pp. 815–830. DOI: 10.1175/jhm-d-17-0205.1. URL: https://journals.ametsoc.org/view/journals/hydr/19/5/jhm-d-17-0205_1.xml.

Baki, H. et al. (2022). "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". In: *Geoscientific Model Development* 15.5, pp. 2133–2155. DOI: 10.5194/gmd-15-2133-2022. URL: https://gmd.copernicus.org/articles/15/2133/2022/.

Bar Massada, A. et al. (2013). "Wildfire ignition-distribution modelling: a comparative study in the Huron–Manistee National Forest, Michigan, USA". In: *International Journal of Wildland Fire* 22.2, pp. 174–183. DOI: 10.1071/WF11178. URL: https://doi.org/10.1071/WF11178.

Bocquet, M. et al. (2020). "Bayesian inference of chaotic dynamics by merging data assimilation, machine learning and expectation-maximization". In: *Foundations of Data Science* 2.1, pp. 55–80. DOI: 10.3934/fods.2020004. URL: /article/id/eb1e67dd-ccb9-4ab2-8526-15c23feae415.

Bonavita, M. and P. Laloyaux (2020). "Machine Learning for Model Error Inference and Correction". In: *Journal of Advances in Modeling Earth Systems* 12.12. e2020MS002232 10.1029/2020MS002232, e2020MS002232. DOI: https://doi.org/10.1029/2020MS002232. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2020MS002232. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020MS002232.

Boser, B. E., I. M. Guyon, and V. N. Vapnik (1992). "A training algorithm for optimal margin classifiers". In: COLT '92, pp. 144–152. DOI: 10.1145/130385.130401. URL: https://doi.org/10.1145/130385.130401.

Box, G. E. and G. C. Tiao (1992). In: *Bayesian Inference in Statistical Analysis*. John Wiley Sons, Ltd. ISBN: 9781118033197. DOI: https://doi.org/10.1002/9781118033197. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118033197. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118033197.

Bratley, P. and B. L. Fox (Mar. 1988). "Algorithm 659: Implementing Sobol's quasirandom sequence generator". In: *ACM Trans. Math. Softw.* 14.1, pp. 88–100. ISSN: 0098-3500. DOI: 10.1145/42288.214372. URL: https://doi.org/10.1145/42288.214372.

https://doi.org/10.5194/gmd-2024-56
Preprint. Discussion started: 18 April 2024

Breiman, L. (2001). "Random Forests". In: *Machine Learning* 45, pp. 5–32. URL: https://api.semanticscholar.org/CorpusID:89141.

Breiman, L. et al. (1984). "Classification and Regression Trees". In: *Biometrics* 40, p. 874. URL: https://api.semanticscholar.org/CorpusID:29458883.

Brunton, S. L. and J. N. Kutz (2019). *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge University Press. DOI: 10.1017/9781108380690.

Campolongo, F., J. Cariboni, and A. Saltelli (2007). "An effective screening design for sensitivity analysis of large models". In: *Environmental Modelling Software* 22.10. Modelling, computer-assisted simulations, and mapping of dangerous phenomena for hazard assessment, pp. 1509–1518. ISSN: 1364-8152. DOI: https://doi.org/10.1016/j.envsoft.2006.10.004. URL: https://www.sciencedirect.com/science/article/pii/S1364815206002805.

Campolongo, F., A. Saltelli, and J. Cariboni (2011). "From screening to quantitative sensitivity analysis. A unified approach". In: *Computer Physics Communications* 182.4, pp. 978–988. ISSN: 0010-4655. DOI: https://doi.org/10.1016/j.cpc.2010.12.039. URL: https://www.sciencedirect.com/science/article/pii/S0010465510005321.

Catani, F. et al. (2013). "Landslide susceptibility estimation by random forests technique: sensitivity and scaling issues". In: *Natural Hazards and Earth System Sciences* 13.11, pp. 2815–2831. DOI: 10.5194/nhess-13-2815-2013. URL: https://nhess.copernicus.org/articles/13/2815/2013/.

Chase, R. J., D. R. Harrison, A. Burke, et al. (2022). "A Machine Learning Tutorial for Operational Meteorology. Part I: Traditional Machine Learning". In: *Weather and Forecasting* 37.8, pp. 1509–1529. DOI: https://doi.org/10.1175/WAF-D-22-0070.1. URL: https://journals.ametsoc.org/view/journals/wefo/37/8/WAF-D-22-0070.1.xml.

Chase, R. J., D. R. Harrison, G. M. Lackmann, et al. (2023). "A Machine Learning Tutorial for Operational Meteorology, Part II: Neural Networks and Deep Learning". In: *Weather and Forecasting*. DOI: https://doi.org/10.1175/WAF-D-22-0187.1. URL: https://journals.ametsoc.org/view/journals/wefo/aop/WAF-D-22-0187.1/WAF-D-22-0187.1.xml.

Chen, T., L. Zhu, et al. (Mar. 2020). "Mapping landslide susceptibility at the Three Gorges Reservoir, China, using gradient boosting decision tree, random forest and information value models". In: *Journal of Mountain Science* 17.3, pp. 670–685. ISSN: 1993-0321. DOI: 10.1007/s11629-019-5839-3. URL: https://doi.org/10.1007/s11629-019-5839-3.

Chen, T. and C. Guestrin (Aug. 2016). "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. ACM. DOI: 10.1145/2939672.2939785. URL: http://dx.doi.org/10.1145/2939672.2939785.

Clark, M. P., D. Kavetski, and F. Fenicia (2011). "Pursuing the method of multiple working hypotheses for hydrological modeling". In: *Water Resources Research* 47.9. DOI: https://doi.org/10.1029/2010WR009827. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2010WR009827. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2010WR009827.

Cortes, C. and V. N. Vapnik (1995). "Support-Vector Networks". In: *Machine Learning* 20, pp. 273–297. URL: https://api.semanticscholar.org/CorpusID:52874011.

Cui, C. and D. Wang (2016). "High dimensional data regression using Lasso model and neural networks with random weights". In: *Information Sciences* 372, pp. 505–517. ISSN: 0020-0255. DOI: https://doi.org/10.1016/j.ins.2016.08.060. URL: https://www.sciencedirect.com/science/article/pii/S0020025516306314.

Cuntz, M., J. Mai, L. Samaniego, et al. (2016). "The impact of standard and hard-coded parameters on the hydrologic fluxes in the Noah-MP land surface model". In: *Journal of Geophysical Research: Atmospheres* 121.18, pp. 10, 676–10, 700. DOI: https://doi.org/10.1002/

https://doi.org/10.5194/gmd-2024-56
Preprint. Discussion started: 18 April 2024

2016JD025097. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2016JD025097. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016JD025097.

Cuntz, M., J. Mai, M. Zink, et al. (2015). "Computationally inexpensive identification of noninformative model parameters by sequential screening". In: *Water Resources Research* 51.8, pp. 6417–6441. DOI: https://doi.org/10.1002/2015WR016907. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2015WR016907. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015WR016907.

Daviran, M. et al. (Jan. 2023). "Landslide susceptibility prediction using artificial neural networks, SVMs and random forest: hyperparameters tuning by genetic optimization algorithm". In: *International Journal of Environmental Science and Technology* 20.1, pp. 259–276. ISSN: 1735-2630. DOI: 10.1007/s13762-022-04491-3. URL: https://doi.org/10.1007/s13762-022-04491-3.

Dey, A. et al. (2022). "A multimodel ensemble machine learning approach for CMIP6 climate model projections in an Indian River basin". In: *International Journal of Climatology* 42.16, pp. 9215–9236. DOI: https://doi.org/10.1002/joc.7813. eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.7813. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.7813.

Di Santo, D. (Mar. 2024). *ML-AMPSIT*. Version v1.0.0. DOI: 10.5281/zenodo.10789930. URL: https://doi.org/10.5281/zenodo.10789930.

Dudhia, J. (1989). "Numerical Study of Convection Observed during the Winter Monsoon Experiment Using a Mesoscale Two-Dimensional Model". In: *Journal of Atmospheric Sciences* 46.20, pp. 3077–3107. DOI: 10.1175/1520-0469(1989)046<3077:NSOCOD>2.0.CO;2. URL: https://journals.ametsoc.org/view/journals/atsc/46/20/1520-0469_1989_046_3077_nsocod_2_0_co_2.xml.

Ek, M. B. et al. (2003). "Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model". In: *Journal of Geophysical Research: Atmospheres* 108.D22. DOI: https://doi.org/10.1029/2002JD003296. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2002JD003296. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2002JD003296.

"Elementary Effects Method" (2007). In: *Global Sensitivity Analysis. The Primer*. John Wiley Sons, Ltd. Chap. 3, pp. 109–154. ISBN: 9780470725184. DOI: https://doi.org/10.1002/9780470725184.ch3. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470725184.ch3. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470725184.ch3.

Elia, L. et al. (July 2023). "Assessing multi-hazard susceptibility to cryospheric hazards: Lesson learnt from an Alaskan example". In: *Science of The Total Environment* 898, p. 165289. DOI: 10.1016/j.scitotenv.2023.165289.

Engelbrecht, A. P., I. Cloete, and J. M. Zurada (1995). *Determining the significance of input parameters using sensitivity analysis*. Undetermined. DOI: 10.1007/3-540-59497-3_199.

Farooq, F. et al. (2020). "A Comparative Study of Random Forest and Genetic Engineering Programming for the Prediction of Compressive Strength of High Strength Concrete (HSC)". In: *Applied Sciences* 10.20. ISSN: 2076-3417. DOI: 10.3390/app10207330. URL: https://www.mdpi.com/2076-3417/10/20/7330.

Fernández-Godino, M. G. et al. (Mar. 2017). "Review of multi-fidelity models". In: *Advances in Computational Science and Engineering* 1, pp. 351–400. DOI: 10.3934/acse.2023015.

Forrester, A., A. Sobester, and A. Keane (July 2008). *Engineering Design Via Surrogate Modelling: A Practical Guide*. ISBN: 978-0-470-06068-1. DOI: 10.1002/9780470770801.

Fowler, H. J., S. Blenkinsop, and C. Tebaldi (2007). "Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling". In: *International Journal of Climatology* 27.12, pp. 1547–1578. DOI: https://doi.org/10.1002/joc.1556. eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.1556. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.1556.

https://doi.org/10.5194/gmd-2024-56
Preprint. Discussion started: 18 April 2024

Gholampour, A., A. H. Gandomi, and T. Ozbakkaloglu (2017). "New formulations for mechanical properties of recycled aggregate concrete using gene expression programming". In: *Construction and Building Materials* 130, pp. 122–145. ISSN: 0950-0618. DOI: https://doi.org/10.1016/j.conbuildmat.2016.10.114. URL: https://www.sciencedirect.com/science/article/pii/S0950061816317408.

Gigović, L. et al. (2019). "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". In: *Forests* 10.5. ISSN: 1999-4907. DOI: 10.3390/f10050408. URL: https://www.mdpi.com/1999-4907/10/5/408.

Grundner, A. et al. (2022). "Deep Learning Based Cloud Cover Parameterization for ICON". In: *Journal of Advances in Modeling Earth Systems* 14.12. e2021MS002959 2021MS002959, e2021MS002959. DOI: https://doi.org/10.1029/2021MS002959. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2021MS002959. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021MS002959.

Haghiabi, A. H., A. H. Nasrolahi, and A. Parsaie (Jan. 2018). "Water quality prediction using machine learning methods". In: *Water Quality Research Journal* 53.1, pp. 3–13. ISSN: 1201-3080. DOI: 10.2166/wqrj.2018.025. eprint: https://iwaponline.com/wqrj/article-pdf/53/1/3/224144/wqrjc0530003.pdf. URL: https://doi.org/10.2166/wqrj.2018.025.

Han, Y. et al. (2020). "A Moist Physics Parameterization Based on Deep Learning". In: *Journal of Advances in Modeling Earth Systems* 12.9. e2020MS002076 2020MS002076, e2020MS002076. DOI: https://doi.org/10.1029/2020MS002076. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2020MS002076. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020MS002076.

Herman, J. D. et al. (2013). "Technical Note: Method of Morris effectively reduces the computational demands of global sensitivity analysis for distributed watershed models". In: *Hydrology and Earth System Sciences* 17.7, pp. 2893–2903. DOI: 10.5194/hess-17-2893-2013. URL: https://hess.copernicus.org/articles/17/2893/2013/.

Hong, S.-Y., Y. Noh, and J. Dudhia (2006). "A New Vertical Diffusion Package with an Explicit Treatment of Entrainment Processes". In: *Monthly Weather Review* 134.9, pp. 2318–2341. DOI: 10.1175/MWR3199.1. URL: https://journals.ametsoc.org/view/journals/mwre/134/9/mwr3199.1.xml.

Kalantar, B. et al. (2018). "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)". In: *Geomatics, Natural Hazards and Risk* 9.1, pp. 49–69. DOI: 10.1080/19475705.2017.1407368. eprint: https://doi.org/10.1080/19475705.2017.1407368. URL: https://doi.org/10.1080/19475705.2017.1407368.

Kim, S. and F. Boukouvala (June 2020). "Machine learning-based surrogate modeling for data-driven optimization: a comparison of subset selection for regression techniques". In: *Optimization Letters* 14. DOI: 10.1007/s11590-019-01428-7.

Kok, Z. H. et al. (2021). "Support Vector Machine in Precision Agriculture: A review". In: *Computers and Electronics in Agriculture* 191, p. 106546. ISSN: 0168-1699. DOI: https://doi.org/10.1016/j.compag.2021.106546. URL: https://www.sciencedirect.com/science/article/pii/S0168169921005639.

Lamberti, G. and C. Gorlé (July 2021). "A multi-fidelity machine learning framework to predict wind loads on buildings". In: *Journal of Wind Engineering and Industrial Aerodynamics* 214, p. 104647. DOI: 10.1016/j.jweia.2021.104647.

Lee, J.-H. et al. (2018). "Modeling landslide susceptibility in data-scarce environments using optimized data mining and statistical methods". In: *Geomorphology* 303, pp. 284–298. ISSN: 0169-555X. DOI: https://doi.org/10.1016/j.geomorph.2017.12.007. URL: https://www.sciencedirect.com/science/article/pii/S0169555X16311060.

Lei, T., S. Ng, and S. Siu (Mar. 2023). "Application of ANN, XGBoost, and Other ML Methods to Forecast Air Quality in Macau". In: *Sustainability* 15, p. 5341. DOI: 10.3390/su15065341.

https://doi.org/10.5194/gmd-2024-56
Preprint. Discussion started: 18 April 2024

Leinonen, J., D. Nerini, and A. Berne (2021). "Stochastic Super-Resolution for Downscaling Time-Evolving Atmospheric Fields With a Generative Adversarial Network". In: *IEEE Transactions on Geoscience and Remote Sensing* 59.9, pp. 7211–7223. DOI: 10.1109/TGRS.2020.3032790.

Li, J. et al. (2018). "Impacts of Land Cover and Soil Texture Uncertainty on Land Model Simulations Over the Central Tibetan Plateau". In: *Journal of Advances in Modeling Earth Systems* 10.9, pp. 2121–2146. DOI: https://doi.org/10.1029/2018MS001377. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2018MS001377. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018MS001377.

Liu, Y. et al. (2021). "A comparative evaluation of machine learning algorithms and an improved optimal model for landslide susceptibility: a case study". In: *Geomatics, Natural Hazards and Risk* 12.1, pp. 1973–2001. DOI: 10.1080/19475705.2021.1955018. eprint: https://doi.org/10.1080/19475705.2021.1955018. URL: https://doi.org/10.1080/19475705.2021.1955018.

Longo, R. et al. (Nov. 2020). "A multi-fidelity framework for the estimation of the turbulent Schmidt number in the simulation of atmospheric dispersion". In: *Building and Environment* 185, p. 107066. DOI: 10.1016/j.buildenv.2020.107066.

Maleki, H. et al. (Aug. 2019). "Air pollution prediction by using an artificial neural network model". In: *Clean Technologies and Environmental Policy* 21.6, pp. 1341–1352. ISSN: 1618-9558. DOI: 10.1007/s10098-019-01709-w. URL: https://doi.org/10.1007/s10098-019-01709-w.

Maraun, D. and M. Widmann (2018). *Statistical Downscaling and Bias Correction for Climate Research*. Cambridge University Press. DOI: 10.1017/9781107588783.

Maroco, J. et al. (Aug. 2011). "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". In: *BMC Research Notes* 4.1, p. 299. ISSN: 1756-0500. DOI: 10.1186/1756-0500-4-299. URL: https://doi.org/10.1186/1756-0500-4-299.

Mckay, M., R. Beckman, and W. Conover (May 1979). "A Comparison of Three Methods for Selecting Vales of Input Variables in the Analysis of Output From a Computer Code". In: *Technometrics* 21, pp. 239–245. DOI: 10.1080/00401706.1979.10489755.

Meenal, R. et al. (Aug. 2022). "Weather Forecasting for Renewable Energy System: A Review". In: *Archives of Computational Methods in Engineering* 29.5, pp. 2875–2891. ISSN: 1886-1784. DOI: 10.1007/s11831-021-09695-3. URL: https://doi.org/10.1007/s11831-021-09695-3.

Mendoza, P. A. et al. (2015). "Are we unnecessarily constraining the agility of complex process-based models?" In: *Water Resources Research* 51.1, pp. 716–728. DOI: https://doi.org/10.1002/2014WR015820. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2014WR015820. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2014WR015820.

Mlawer, E. J. et al. (1997). "Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave". In: *Journal of Geophysical Research: Atmospheres* 102.D14, pp. 16663–16682. DOI: https://doi.org/10.1029/97JD00237. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/97JD00237. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/97JD00237.

Mooers, G. et al. (2021). "Assessing the Potential of Deep Learning for Emulating Cloud Superparameterization in Climate Models With Real-Geography Boundary Conditions". In: *Journal of Advances in Modeling Earth Systems* 13.5. e2020MS002385 2020MS002385, e2020MS002385. DOI: https://doi.org/10.1029/2020MS002385. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2020MS002385. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020MS002385.

Morris, M. D. (1991). "Factorial sampling plans for preliminary computational experiments". In: *Quality Engineering* 37, pp. 307–310.

https://doi.org/10.5194/gmd-2024-56
Preprint. Discussion started: 18 April 2024

Murti, M. A. et al. (Dec. 2022). "Earthquake multi-classification detection based velocity and displacement data filtering using machine learning algorithms". In: *Scientific Reports* 12.1, p. 21200. ISSN: 2045-2322. DOI: 10.1038/s41598-022-25098-1. URL: https://doi.org/10.1038/s41598-022-25098-1.

Muthukrishnan, R. and R. Rohini (2016). "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. DOI: 10.1109/ICACA.2016.7887916.

Niu, G.-Y. and Z.-L. Yang (2004). "Effects of vegetation canopy processes on snow surface energy and mass balances". In: *Journal of Geophysical Research: Atmospheres* 109.D23. DOI: https://doi.org/10.1029/2004JD004884. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2004JD004884. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2004JD004884.

– (2007). "An observation-based formulation of snow cover fraction and its evaluation over large North American river basins". In: *Journal of Geophysical Research: Atmospheres* 112.D21. DOI: https://doi.org/10.1029/2007JD008674. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2007JD008674. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2007JD008674.

Niu, G.-Y., Z.-L. Yang, et al. (2011). "The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements". In: *Journal of Geophysical Research: Atmospheres* 116.D12. DOI: https://doi.org/10.1029/2010JD015139. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2010JD015139. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2010JD015139.

O'Hagan, A. (2006). "Bayesian analysis of computer code outputs: A tutorial". In: *Reliability Engineering System Safety* 91.10. The Fourth International Conference on Sensitivity Analysis of Model Output (SAMO 2004), pp. 1290–1300. ISSN: 0951-8320. DOI: https://doi.org/10.1016/j.ress.2005.11.025. URL: https://www.sciencedirect.com/science/article/pii/S0951832005002383.

Oliveira, S. et al. (2012). "Modeling spatial patterns of fire occurrence in Mediterranean Europe using Multiple Regression and Random Forest". In: *Forest Ecology and Management* 275, pp. 117–129. ISSN: 0378-1127. DOI: https://doi.org/10.1016/j.foreco.2012.03.003. URL: https://www.sciencedirect.com/science/article/pii/S0378112712001272.

Palani, S., S.-Y. Liong, and P. Tkalich (2008). "An ANN application for water quality forecasting". In: *Marine Pollution Bulletin* 56.9, pp. 1586–1597. ISSN: 0025-326X. DOI: https://doi.org/10.1016/j.marpolbul.2008.05.021. URL: https://www.sciencedirect.com/science/article/pii/S0025326X08003196.

Pourtaghi, Z. S. et al. (2016). "Investigation of general indicators influencing on forest fire and its susceptibility modeling using different data mining techniques". In: *Ecological Indicators* 64, pp. 72–84. ISSN: 1470-160X. DOI: https://doi.org/10.1016/j.ecolind.2015.12.030. URL: https://www.sciencedirect.com/science/article/pii/S1470160X15007359.

Pradhan, B. (2013). "A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS". In: *Computers Geosciences* 51, pp. 350–365. ISSN: 0098-3004. DOI: https://doi.org/10.1016/j.cageo.2012.08.023. URL: https://www.sciencedirect.com/science/article/pii/S0098300412003093.

Queipo, N. V. et al. (2005). "Surrogate-based analysis and optimization". In: *Progress in Aerospace Sciences* 41.1, pp. 1–28. ISSN: 0376-0421. DOI: https://doi.org/10.1016/j.paerosci.2005.02.001. URL: https://www.sciencedirect.com/science/article/pii/S0376042105000102.

Rahmati, O., H. R. Pourghasemi, and A. M. Melesse (2016). "Application of GIS-based data driven random forest and maximum entropy models for groundwater potential mapping: A case study at Mehran Region, Iran". In: *CATENA* 137, pp. 360–372. ISSN: 0341-8162. DOI: https://doi.org/10.1016/j.catena.2015.10.010. URL: https://www.sciencedirect.com/science/article/pii/S0341816215301326.

Rasmussen, C. E. and C. K. I. Williams (Nov. 2005). *Gaussian Processes for Machine Learning*. The MIT Press. ISBN: 9780262256834. DOI: 10.7551/mitpress/3206.001.0001. URL: https://doi.org/10.7551/mitpress/3206.001.0001.

Rasp, S., M. S. Pritchard, and P. Gentine (2018). "Deep learning to represent subgrid processes in climate models". In: *Proceedings of the National Academy of Sciences* 115.39, pp. 9684–9689. DOI: 10.1073/pnas.1810286115. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1810286115. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1810286115.

Ren, X. et al. (Dec. 2020). "Deep Learning-Based Weather Prediction: A Survey". In: *Big Data Research* 23, p. 100178. DOI: 10.1016/j.bdr.2020.100178.

Ren, Y., L. Zhang, and P. Suganthan (2016). "Ensemble Classification and Regression-Recent Developments, Applications and Future Directions [Review Article]". In: *IEEE Computational Intelligence Magazine* 11.1, pp. 41–53. DOI: 10.1109/MCI.2015.2471235.

Rodriguez-Galiano, V. et al. (2014). "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)". In: *Science of The Total Environment* 476-477, pp. 189–206. ISSN: 0048-9697. DOI: https://doi.org/10.1016/j.scitotenv.2014.01.001. URL: https://www.sciencedirect.com/science/article/pii/S0048969714000102.

Ross, A. et al. (2023). "Benchmarking of Machine Learning Ocean Subgrid Parameterizations in an Idealized Model". In: *Journal of Advances in Modeling Earth Systems* 15.1. e2022MS003258 2022MS003258, e2022MS003258. DOI: https://doi.org/10.1029/2022MS003258. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2022MS003258. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2022MS003258.

Sakaguchi, K. and X. Zeng (2009). "Effects of soil wetness, plant litter, and under-canopy atmospheric stability on ground evaporation in the Community Land Model (CLM3.5)". In: *Journal of Geophysical Research: Atmospheres* 114.D1. DOI: https://doi.org/10.1029/2008JD010834. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2008JD010834. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2008JD010834.

Salmasi, F. et al. (Sept. 2020). "Application of SVM, ANN, GRNN, RF, GP and RT models for predicting discharge coefficients of oblique sluice gates using experimental data". In: *Water Supply* 21.1, pp. 232–248. ISSN: 1606-9749. DOI: 10.2166/ws.2020.226. eprint: https://iwaponline.com/ws/article-pdf/21/1/232/839979/ws021010232.pdf. URL: https://doi.org/10.2166/ws.2020.226.

Saltelli, A., P. Annoni, et al. (2010). "Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index". In: *Computer Physics Communications* 181.2, pp. 259–270. ISSN: 0010-4655. DOI: https://doi.org/10.1016/j.cpc.2009.09.018. URL: https://www.sciencedirect.com/science/article/pii/S0010465509003087.

Saltelli, A., M. Ratto, et al. (2008). *Global Sensitivity Analysis: The Primer*. URL: https://api.semanticscholar.org/CorpusID:115957810.

Saltelli, A. and I. Sobol' (Jan. 1995). "Sensitivity analysis for nonlinear mathematical models: Numerical experience". In: *Matematicheskoe Modelirovanie* 7.

Sarkar, A. and P. Pandey (2015). "River Water Quality Modelling Using Artificial Neural Network Technique". In: *Aquatic Procedia* 4. INTERNATIONAL CONFERENCE ON WATER RESOURCES, COASTAL AND OCEAN ENGINEERING (ICWRCOE'15), pp. 1070–1077. ISSN: 2214-241X. DOI: https://doi.org/10.1016/j.aqpro.2015.02.135. URL: https://www.sciencedirect.com/science/article/pii/S2214241X15001364.

Satir, O., S. Berberoglu, and C. Donmez (2016). "Mapping regional forest fire probability using artificial neural network model in a Mediterranean forest ecosystem". In: *Geomatics, Natural Hazards and Risk* 7.5, pp. 1645–1658. DOI: 10.1080/19475705.2015.1084541. eprint: https://doi.org/10.1080/19475705.2015.1084541. URL: https://doi.org/10.1080/19475705.2015.1084541.

Schaake, J. C. et al. (1996). "Simple water balance model for estimating runoff at different spatial and temporal scales". In: *Journal of Geophysical Research: Atmospheres* 101.D3, pp. 7461–7475. DOI: https://doi.org/10.1029/95JD02892. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/95JD02892. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/95JD02892.

https://doi.org/10.5194/gmd-2024-56
Preprint. Discussion started: 18 April 2024

Schneider, R. et al. (June 2022). "ESA-ECMWF Report on recent progress and research directions in machine learning for Earth System observation and prediction". In: *npj Climate and Atmospheric Science* 5.1, p. 51. ISSN: 2397-3722. DOI: 10.1038/s41612-022-00269-z. URL: https://doi.org/10.1038/s41612-022-00269-z.

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

Schultz, M. G. et al. (2021). "Can deep learning beat numerical weather prediction?" In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 379.2194, p. 20200097. DOI: 10.1098/rsta.2020.0097. eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rsta.2020.0097. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2020.0097.

Sekulić, A. et al. (2020). "Random Forest Spatial Interpolation". In: *Remote Sensing* 12.10. ISSN: 2072-4292. DOI: 10.3390/rs12101687. URL: https://www.mdpi.com/2072-4292/12/10/1687.

Shah, M. I. et al. (2021). "Predictive Modeling Approach for Surface Water Quality: Development and Comparison of Machine Learning Models". In: *Sustainability* 13.14. ISSN: 2071-1050. DOI: 10.3390/su13147515. URL: https://www.mdpi.com/2071-1050/13/14/7515.

Sharma, A. et al. (Apr. 2014). "A feature selection method using improved regularized linear discriminant analysis". In: *Machine Vision and Applications* 25.3, pp. 775–786. ISSN: 1432-1769. DOI: 10.1007/s00138-013-0577-y. URL: https://doi.org/10.1007/s00138-013-0577-y.

Shen, K.-Q. et al. (Jan. 2008). "Feature selection via sensitivity analysis of SVM probabilistic outputs". In: *Machine Learning* 70.1, pp. 1–20. ISSN: 1573-0565. DOI: 10.1007/s10994-007-5025-7. URL: https://doi.org/10.1007/s10994-007-5025-7.

Sihag, P. et al. (Dec. 2019). "Predictive modeling of PM2.5 using soft computing techniques: case study—Faridabad, Haryana, India". In: *Air Quality, Atmosphere & Health* 12.12, pp. 1511–1520. ISSN: 1873-9326. DOI: 10.1007/s11869-019-00755-z. URL: https://doi.org/10.1007/s11869-019-00755-z.

Skamarock, W. C. et al. (2021). "A Description of the Advanced Research WRF Model Version 4.3". In: *No. NCAR/TN-556+STR*. DOI: 10.5065/1dfh-6p97.

Sridhara, S. et al. (Jan. 2023). "Evaluation of machine learning approaches for prediction of pigeon pea yield based on weather parameters in India". In: *International Journal of Biometeorology* 67.1, pp. 165–180. ISSN: 1432-1254. DOI: 10.1007/s00484-022-02396-x. URL: https://doi.org/10.1007/s00484-022-02396-x.

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

Suárez Sánchez, A. et al. (2011). "Application of an SVM-based regression model to the air quality study at local scale in the Avilés urban area (Spain)". In: *Mathematical and Computer Modelling* 54.5, pp. 1453–1466. ISSN: 0895-7177. DOI: https://doi.org/10.1016/j.mcm.2011.04.017. URL: https://www.sciencedirect.com/science/article/pii/S0895717711002196.

Tian, W. (2013). "A review of sensitivity analysis methods in building energy analysis". In: *Renewable and Sustainable Energy Reviews* 20, pp. 411–419. ISSN: 1364-0321. DOI: https://doi.org/10.1016/j.rser.2012.12.014. URL: https://www.sciencedirect.com/science/article/pii/S1364032112007101.

Tibshirani, R. (1996). "Regression Shrinkage and Selection Via the Lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1, pp. 267–288. DOI: https://doi.org/10.1111/j.2517-6161.1996.tb02080.x. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1996.tb02080.x. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1996.tb02080.x.

Torres, M. (June 2021). "A Machine Learning Method for Parameter Estimation and Sensitivity Analysis". In: pp. 330–343. ISBN: 978-3-030-77976-4. DOI: 10.1007/978-3-030-77977-1_26.

https://doi.org/10.5194/gmd-2024-56
Preprint. Discussion started: 18 April 2024

Trabelsi, F. and S. Bel Hadj Ali (2022). "Exploring Machine Learning Models in Predicting Irrigation Groundwater Quality Indices for Effective Decision Making in Medjerda River Basin, Tunisia". In: *Sustainability* 14.4. ISSN: 2071-1050. DOI: 10.3390/su14042341. URL: https://www.mdpi.com/2071-1050/14/4/2341.

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

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

Wolff, B., O. Kramer, and D. Heinemann (2017). "Selection of Numerical Weather Forecast Features for PV Power Predictions with Random Forests". In: *Data Analytics for Renewable Energy Integration*. Ed. by W. L. Woon et al. Cham: Springer International Publishing, pp. 78–91. ISBN: 978-3-319-50947-1.

Wu, W. et al. (Oct. 2021). "The transferability of random forest and support vector machine for estimating daily global solar radiation using sunshine duration over different climate zones". In: *Theoretical and Applied Climatology* 146.1, pp. 45–55. ISSN: 1434-4483. DOI: 10.1007/s00704-021-03726-6. URL: https://doi.org/10.1007/s00704-021-03726-6.

Wu, Z. et al. (2023). "Simulation of daily maize evapotranspiration at different growth stages using four machine learning models in semi-humid regions of northwest China". In: *Journal of Hydrology* 617, p. 128947. ISSN: 0022-1694. DOI: https://doi.org/10.1016/j.jhydrol.2022.128947. URL: https://www.sciencedirect.com/science/article/pii/S0022169422015177.

Yang, W.-X. et al. (Dec. 2022). "Comparison of ischemic stroke diagnosis models based on machine learning". In: *Frontiers in Neurology* 13. DOI: 10.3389/fneur.2022.1014346.

Yang, Z.-L. et al. (2011). "The community Noah land surface model with multiparameterization options (Noah-MP): 2. Evaluation over global river basins". In: *Journal of Geophysical Research: Atmospheres* 116.D12. DOI: https://doi.org/10.1029/2010JD015140. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2010JD015140. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2010JD015140.

Yilmaz, I. (Aug. 2010). "Comparison of landslide susceptibility mapping methodologies for Koyulhisar, Turkey: conditional probability, logistic regression, artificial neural networks, and support vector machine". In: *Environmental Earth Sciences* 61.4, pp. 821–836. ISSN: 1866-6299. DOI: 10.1007/s12665-009-0394-9. URL: https://doi.org/10.1007/s12665-009-0394-9.

Youssef, A. M. et al. (Oct. 2016). "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". In: *Landslides* 13.5, pp. 839–856. ISSN: 1612-5118. DOI: 10.1007/s10346-015-0614-1. URL: https://doi.org/10.1007/s10346-015-0614-1.

Yu, R. et al. (2016). "RAQ–A Random Forest Approach for Predicting Air Quality in Urban Sensing Systems". In: *Sensors* 16.1. ISSN: 1424-8220. DOI: 10.3390/s16010086. URL: https://www.mdpi.com/1424-8220/16/1/86.

Yuval, J. and P. A. O'Gorman (July 2020). "Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions". In: *Nature Communications* 11.1, p. 3295. ISSN: 2041-1723. DOI: 10.1038/s41467-020-17142-3. URL: https://doi.org/10.1038/s41467-020-17142-3.

Zhang, X., R. Srinivasan, and M. Van Liew (2009). "Approximating SWAT Model Using Artificial Neural Network and Support Vector Machine1". In: *JAWRA Journal of the American Water Resources Association* 45.2, pp. 460–474. DOI: https://doi.org/10.1111/j.1752-1688.2009.00302.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1752-1688.2009.00302.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1752-1688.2009.00302.x.

https://doi.org/10.5194/gmd-2024-56
Preprint. Discussion started: 18 April 2024

Zhou, C. et al. (2018). "Landslide susceptibility modeling applying machine learning methods: A case study from Longju in the Three Gorges Reservoir area, China". In: *Computers Geosciences* 112, pp. 23–37. ISSN: 0098-3004. DOI: https://doi.org/10.1016/j.cageo.2017.11.019. URL: https://www.sciencedirect.com/science/article/pii/S0098300417304843.

Zouhri, W., L. Homri, and J.-Y. Dantan (2022). "Handling the impact of feature uncertainties on SVM: A robust approach based on Sobol sensitivity analysis". In: *Expert Systems with Applications* 189, p. 115691. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2021.115691. URL: https://www.sciencedirect.com/science/article/pii/S0957417421010769.