https://doi.org/10.5194/egusphere-2022-465
Preprint. Discussion started: 5 July 2022

UVBoost (v0.5): a hybrid radiative transfer and machine learning model for estimating ultraviolet radiation

Marcelo de Paula Corrêa[1]

[1]Instituto de Recursos Naturais, Universidade Federal de Itajubá, Itajubá/MG, 37500-903, Brazil

Correspondence to: Marcelo de Paula Corrêa (mpcorrea@unifei.edu.br)

Abstract. This article presents UVBoost, a hybrid radiative transfer estimator based on a Supervised Machine Learning (SML) regression model powered by high precision ultraviolet radiation (UVR) calculations provided by a conventional Radiative Transference Model (RTM). The proposed regression model takes UVR as a dependent variable, and the Solar Zenith Angle (SZA), Total Ozone Content (TOC), and Aerosol Optical Depth (AOD), as the independent predictive variables. UVBoost was developed to increase computational speed for conducting calculations with large databases, without sacrificing result accuracy. Furthermore, this method employs a user-friendly code, which can be used by laymen or researchers in other areas. UVBoost can be used to disseminate UVR data online anywhere in different spatiotemporal scales, or for climatological projection studies on a global scale. The model was developed by comparing seven regression SML tools via cross validation. These results were validated using non-parametric statistical tests. Of all the tested tools, the Categorical Boosting (CatBoost) method showed the best accuracy at the lowest computational cost. Two additional studies were carried out, one at the global scale, and another at the local scale, to compare the traditional RTM vs. the UVBoost results. The first study simulated a global UVR field (1°x1°), with 64800 grid points, with input data from CMIP6, available at https://pcmdi.llnl.gov/CMIP6/. The differences between the RTM and the UVBoost were less than ±5% for approximately 95% of all points, except for points with high SZA. The computational speed of UVBoost surpassed that of the RTM by more than three orders of magnitude. The second study simulated the daily UVR at eight different locations on Earth. The results showed that the UVBoost was very efficient in simulating accumulated UVR doses during the day, with negligible differences (< ±3%), which means it can be used in studies on UVR and human health. In the future, UVBoost will include other geophysical parameters and be extended to other bands in the electromagnetic spectrum.

1 Introduction

Ultraviolet radiation (UVR) causes important photochemical and photobiological effects in the atmosphere and surface of the Earth. In the upper atmosphere, UVR plays a key role in the stratosphere in ozone production and destruction processes, known as the Chapman cycle (Chapman, 1930). In low levels, UVR accelerates the reaction between volatile organic compounds (VOC) and oxides and nitrogen (NOx) on the surface, to form ozone molecules, which are important secondary pollutants in large cities (Finlayson-Pitts and Pitts, 2000). UVR causes acute photobiological effects on humans e.g., Erythema, sun burns, and eye inflammations, like photokeratitis and photoconjunctivitis. Long-term cumulative effects include premature aging, damage to hair, skin cancer, etc. (ICNIRP, 2004; DeVecchi et al., 2019). By contrast, UVR exposure is also associated with several beneficial effects, including synthesizing vitamin D, and preventing several diseases, like some types of cancers and diabetes (Gorman et al., 2017). Beyond the human scope, UVR causes photobiological effects for all living beings, including microorganisms, birds, mammals, and plants (Schmalwieser, 2018).

Skin cancer cases have increased worldwide from 3 to 7% over the last decade (Leiter et al., 2020). Disseminating information on UVR levels is a practice that is recommended by the World Health Organization (WHO) to mitigate this problem, and also to develop efficient public policies for photoprotection (WHO, 2002). UVR measurements are essential if these actions are to be viable, and are essential for studies on any of the aforementioned effects, especially those related to human health. However, in situ UVR data are scarce and poorly temporally and spatially distributed worldwide (Liu et al., 2017). Remote sensing is an alternative way of conducting measurements. However, UVR satellite measurements have significant limitations and uncertainties, especially for sporadic measurements and under certain atmospheric conditions where errors can be greater than 50% (Jégou et al., 2011; Wenmin et al., 2020).

Thus, Radiative Transference Models (RTM) have been widely applied to assess and predict UVR over large space-time scales, and in studies on climate projections. Together with in situ or remote sensing measurements, RTMs are essential for monitoring and studying UVR. The most widely used RTMs for research on UVR are LibRadTran (Emde et al., 2016), SBDART (Ricchiazzi et al., 1998), TUV (Madronich and Flocke, 1997), which are very accurate computational codes estimating UVR, especially under clear-sky conditions, or in the presence of aerosols and stratified clouds. Despite showing a good relationship between performance, accuracy, and computational costs, RTMs can require long calculation times under certain situations, e.g., highly detailed grid point global assessments, or for time series climatological data. Another problem related to RTM is that these mathematical models are complex, and require technical and computational knowledge to install and operate them, with a few exceptions e.g., Quick TUV calculator - https://www.acom.ucar.edu/Models/TUV/Interactive_TUV/.

Recent advancements in different areas of knowledge, and the popularization of Supervised Machine Learning (SML) techniques applied to modeling via regression and classification, have opened up a new field for solving problems in the Atmospheric Sciences. SML models in conjunction with robust and accurate databases, allows researchers to prediction extreme situations, or extract events quickly and accurately. Thus, hybrid modeling, which merges physical process models and SML, is now essential in improving seasonal forecasting and modeling at various scales of time and space (Reichstein et al., 2019).

Given these innovative techniques, this paper presents a hybrid model for UVR calculations, which is called the “Boosted Hybrid UV Radiative Transference Model (UVBoost)”. UVBoost uses a high-resolution database provided by a RTM combined with an SML regression model. UVBoost was tested against other common SML regression techniques, including traditional multiple linear regression models and other different improved methods using decision tree models. The results showed that the CatBoost (Hancock and Khoshgoftaar, 2020) tool was a powerful and fast algorithm that uses decision trees.

UVBoost estimates UVR fluxes at different Solar Zenith Angles (SZA), Total Ozone Content (TOC), and Aerosol Optical Depths (AOD), observed at different times of the year, across the entire planet, and does so very quickly with very high precision.

2 Methods and Model Description

2.1 Stochastic SML Models

Multiple Linear Regression (MLR) is a technique used when there are different predictors (X_1, X_2, ..., X_n) in determining any dependent variable (Y). For example, Eq. (1) is an MLR with the same variables used in this study, where b_1, b_2, ..., b_n are coefficients (weights) for each predictor variable, and b_0 is the Y intercept.

$$UVR = b_0 + b_1 SZA + b_2 TOC + b_3 AOD \qquad (1)$$

Generally, even if these predictors are statistically significant, the MLR still has many limitations. The predicted Y value can vary as a function of the coefficients b_j for each unit of change in X_j, assuming that all other variables, X_k to $k \neq j$, remain the same. Thus, for this study, the different orders of magnitude, linearity, and variable increments, along with other variables, limit MLR use as a predictive UVR model.

Cross-validation is a widely employed technique for minimizing classic sample-based statistical regression metric errors. This technique divides the dataset into multiple random samples to train and test the model. The most common technique is k-fold cross-validation, which reserves a 1/k data portion as a test sample, while the rest of the data is used for training. The coefficients obtained in the training model are applied to a test where the evaluation metrics are recorded. Then, the next 1/k in the data, which was not used in the previous sample, is selected, and the previous 1/k group is replaced. This is repeated until each element in the dataset has been used in the sample test. The average of, or a combination of the evaluation metrics should be as an optimized result. However, even when using cross-validation techniques one cannot properly fit an MLR model to UVR predictions.

Stochastic SML models do not depend on linearity to statistically treat data, unlike traditional statistical techniques. One of the first stochastic SML models was a Support Vector Machine (SVM). The original algorithms were developed in the 1960s. SVMs are a set of supervised learning methods that are used for classifications, regressions, and for detecting outliers. They use a subset of points for training in the decision function, which are called support vectors. These models are versatile and effective over large spaces, even when the number of dimensions is greater than the number of samples. Furthermore, SVMs are not influenced by outliers, and can be applied to solve linear and non-linear problems. However, this model has problems when graphically visualizing and theoretically interpreting results, given the complex mathematics. Furthermore, it is a slow model compared to other algorithms. Lastly, parameters must be carefully adjusted to prevent overfitting and underfitting.

Decision tree models (DT), which were developed in the mid-1980s, have proved to be more efficient than SVMs. DTs have become so popular that they created an entirely new domain for developing many descending techniques, e.g., random forests (RF), or different boosting techniques. These innovations have led to better more powerful predictions, and now form the base for other predictive models used in data science.

DT models are sets of *if-then-else* rules that can determine patterns in complex iterations in large data sets, that take Y as a dependent response variable, for a C set of independent predictor variables X_i, where i = 1, 2, ..., C. For a k record partition, recursive partitioning will find the best way of dividing k into two sub-partitions. For each a_i independent variable value (X_i), k with X_i values are divided into one group with values that are smaller than a_i, and the remaining ($X_i \geq a_i$) are set into another group. The homogeneity of the data classes within each k sub-group is measured, and the a_i value resulting in the maximum class homogeneity is selected from among the sets. This must occur recursively, where k is initialized with the entire dataset. The partitioning algorithm divides k into groups k_1 and k_2, and this is repeated, hence the tree branches, for the two groups, until homogeneity for these sub-allocations cannot be increased. Generally speaking, homogeneity, or class purity, is measured using breakdown accuracy coefficients. Precision is represented as a proportion (p) of wrongly classified records within a given group, ranging from 0 (accurate) to 0.5 (completely random). The most widely used coefficients are the Gini impurity coefficient $I(k) = p(1 - p)$, and the entropy coefficient, $I(k) = - p\,log_2(p) - (1 - p)log_2(1 - p)$. Both coefficients give similar results, but the entropy coefficient results in higher impurity values at higher accuracy levels.

Generally, these decision trees need to be 'pruned' to prevent very small groups from forming at the ends. The partitioning process stops when sub-allocations are too small, and this is usually determined via arbitrary rules. It can also be stopped when re-partitions do not significantly reduce impurity. In the latter case, the complexity parameter (cp) is used to estimate the tree size that results in the best performance given the database. Very large cp values reduce predictive capacity by producing very small trees. By contrast, very small cp leads to overfitting. Generally, cross-validation is used to determine the best biases and data variance relationships to find the best cp value. The data set is divided into both training and testing groups. The training group trains the decision tree, and pruning is performed successively, with consequent cp records, which correspond to minimum data validation errors. The division into training/testing groups, and growing, pruning, and recording the cp is repeated to determine the average cp that results in minimum errors for each DT. Once the best cp (minimum error) has been determined, a return function goes back to the original data set to build a DT that will be pruned according to this particular cp.

ML predictive models can be improved by using a model clustering technique, i.e., the average of the results from multiple models, which tends to be more accurate, and tends to have less bias than from a single model. The aggregating bootstrap (bagging) technique, which is quite common in ML models, fits each model from the bootstrap sample instead of conventionally fitting it with different models using the same data, e.g., returning to the Y answer and to the i predictor variables (X_1, X_2,..., Xi). Assuming N models for fitting, n (n < N) training records are then selected. We start from the first iteration (n = 1), and remove the bootstrap sample by replacing the n training data records to form the first subsample Y_N to

X_N. Then, the model is trained using Y_N and X_N to create a set of decision rules $r_n(X)$. It then proceeds to the next iteration (m = 2) until n = N. Finally, the decision is estimated using $r = \frac{1}{N}(r_1(X) + r_2(X) + \cdots + r_N(X))$.

The RF method applies bagging to RTM with samples from both the records and the variables, i.e., the chosen variable is limited to a random group of variables at each stage of the algorithm, and includes bootstrap sampling by replacing variables in each division, in addition to bagging. First, a bootstrap sample is taken from the records. In the first division, i < I variables are randomly sampled without replacements. The division algorithm is applied for each of these sampled variables (X_{j1}, X_{j2},..., X_{ji}). Thus, for each a_{jk} of X_{jk} value, the records in group A are divided, with $a_{jk} > X_{jk}$ as a partition, while the remaining records $a_{jk} \leq X_{jk}$ are taken as another partition. The homogeneity of the classes within each partition is measured. The partition values and variable a_{jk} and X_{jk} that result in maximum homogeneity are selected. The next step is the subsequent partition, and these same steps are repeated until the tree has grown sufficiently. The whole process is repeated with the new bootstrap sample. RF contains a set of hyperparameters that need to be adjusted via cross-validation to prevent overfitting. This is a "black box" method, and produces more accurate predictions than DT, albeit by sacrificing intuitive characteristics.

These clustered models are actually standard in predictive models. Boosting is another widely used technique for forming model clusters, which used residual evaluations to improve fits. This technique is similar to bagging, but it is more refined. Basically, iterations start using the maximum number (N) of models for fitting. The iterations have observational weights w_i = 1/N, where i = 1, 2, ..., N, starting with the G_0 = 0 clustering model. The model is trained using r_n, with the observational weights w_1, w_2 , ..., w_n, which decrease the weighted error e_n defined by the sum of the weights of the incorrect classifications. This model is added to the $G_n = G_{n-1} + \beta_n r_n$ group, where $\beta_n = \frac{\log(1-e_n)}{e_n}$. The w_i weights need to be updated at each step, and are increased for misclassified observations. This increase results in heavier trainings for worse performing data. The degree of increase in w_i depends on β_n, which is higher the greater the β_n. These iterations need to be performed until n = N. In the end, the boosted estimate is expressed by $G = \beta_1 r_1 + \beta_2 r_2 + \cdots + \beta_n r_n$. This type of iteration is the foundation of boosting models, but several current variations use cost function optimizations, randomness, and model adjustments with residuals, for example.

Some improved gradient boosting methods are currently widely employed, e.g., XGBoost (Chen and Guestrin, 2016), LightGBM (Ke et al., 2017), and CatBoost (Hancock and Khoshgoftaar, 2020). These are common in the scientific community, and are available in several programming languages. They are fast and have a wide range of hyperparameters. Thus, cross-validation is practically required for obtaining the best fit for these models. XGBoost works with sparse data, and uses cache access patterns, data compression, and cluster fragmentation to build a fast and flexible DT growth system, at minimal computational resource costs. LightGBM, by contrast, is faster than XGBoost. It features gradient-based sampling and unique feature clustering to handle a large data sets and resource instances. Finally, the more recent CatBoost (Categorial Boosting), creates a continuous DT set. One of the main differences between CatBoost and other algorithms is that the partition criterion is used at all tree levels (symmetric trees). Therefore, calculations are less prone to overfitting, and execution time is significantly sped up. CatBoost performs gradient boosting using ordered boosting, i.e., if the data does not stipulate a time range, CatBoost randomly creates an artificial time point for each data point. The residual is calculated for each datapoint using the most current training from the other datapoints. Different models are trained to calculate the following residuals for other datapoints using data that have not been input to the model. This is repeated for each iteration. CatBoost divides the database into random permutations to apply ordered boosting. This randomness prevents overfitting, resulting in it being more accurate.

2.2 Building a database to train and test the UVBoost model

Regression models using SML are powerful tools since they can be built from known and validated assumptions (Reichstein et al., 2019). The database must be sufficiently consistent and devoid of missing, inconsistent, or conflicting data, to prevent error propagation (Braiek and Khomh, 2020). Furthermore, tests for building the ML model must consider other failure sources, e.g., techniques that explore cross-validations between the testing sets and training sets.

The database for testing and training, in this study, was the RTM TUV v5.3.2 (Madronich and Flocke, 1997), available at https://www2.acom.ucar.edu/modeling/tuv-download. The TUV is a one-dimensional multilayer model that solves the radiative transference equation using a two-flow method, or using discrete ordinates (DISORT) with *n* fluctuations. The UVR irradiances were calculated using high precision parameterization, with discrete ordinates at 8 streams, which allows for more accurate estimates, albeit, at the expense of high computational costs. The database for testing the SML models was designed and built to avoid poor quality information errors. UVR were estimated based on the SZA, TOC and AOD input parameters, in addition to the high precision calculations, for climatological conditions in different regions of the planet. I considered possible future variability for these parameters, as predicted in the Coupled Model Intercomparison Project Phase 6 (CMIP6) (Eyring et al., 2016) (Table 1), to guarantee the applicability of the UVBoost under all atmospheric conditions.

Other atmospheric parameters that exert less influence on the surface fluctuations were kept constant, e.g., the vertical structure of the atmosphere, or the presence of other atmospheric components that do not affect UVR. It is worth mentioning that cloud presence was not considered in this study for two reasons: a) Large variability and spatio-temporal complexity for planet cloud cover means less accurate RTM results; and, b) The Cloud Modification Factor (CMF) is widely used to estimate UVR fluctuations under cloudy conditions (Blumthaler, 2018; Vuilleumier et al., 2021).

The CMF is a multiplicative factor that represents cloud cover. It is used in conjunction with other UVR calculations using RTM applied to cloudless skies (Foyo-Moreno et al, 2001). Thus, cloudiness is directly related to good estimates for clear skies, which is precisely the scope of this study.

The model output was adjusted to generate irradiances, in the absence of cloudiness, weighted by the photobiological responses to the harmful effects of UVR, e.g., erythema (UVE) (ISO/CIE, 1999). For practicality's sake and for didactic purposes, UVE fluctuations were represented using the Ultraviolet Index (UVI). The UVI is an easy-to-understand dimensionless scale for public use. It is also used by the World Health Organization (WMO) to disseminate UVR fluxes that could be harmful to human health (WHO, 2002). One UVI unit is equivalent to 0.025 Wm^{-2} of UVE. Tests were also performed with UVR weighted with photobiological responses for vitamin D synthesis (UVD) (CIE, 2006) with very similar results to UVI results. For simplicity's sake, the results in this article will focus on UVI and erythemal doses.

Table 1: RTM TUV Input parameters

Solar Zenith Angle (SZA)	0° to 90°, step = 1.0°
Total Ozone Content (TOC)	0 to 650 DU, step = 10 DU
Aerosol Optical Depth (AOD)	0.0 to 2.0, step = 0.1; from 2.0 to 5.0, step = 0.5; and, from 5.0 to 15.0, step = 2.5.

Figure 1 shows UVR sensitivity relative to input parameter variations, as shown in Table 1. Figures 1a and 1b, show the UVI variation for TOC and AOD variations at different sun positions (SZA), respectively. One can see that high UVI levels correspond to lower zenith angles, less ozone and fewer aerosols in the atmosphere. This explains the high UVI levels at the equator and in subequatorial regions, where TOC is in the order of 250 to 280 DU, where AOD is less than 0.1, and where SZA values are less than 20° close to sun solar noon. Another aspect worth noting is that atmospheric aerosols have a minor influence on UVR fluxes relative to TOC. AODs less than 0.2 were observed for most parts of the planet, all throughout the year. By contrast, AOD values > 0.2 were sporadic, and were generally observed for highly polluted regions over large cities (e.g., Beijing, China; Santiago, Chile; etc.) or over desert regions in Central Africa or the Middle East (Sogacheva et al., 2020). I also analysed variations in other aerosol optical properties, e.g., the simple albedo, the asymmetry parameter, and the Ånsgtröm coefficient, but the influence of these parameters on UVR fluctuations was negligible relative to amplitudes commonly observed at different regions. on Earth. Furthermore, these parameters do not comprise climatological databases like the CMIP5 or CMIP6, for example, and therefore were irrelevant within the scope of this study.

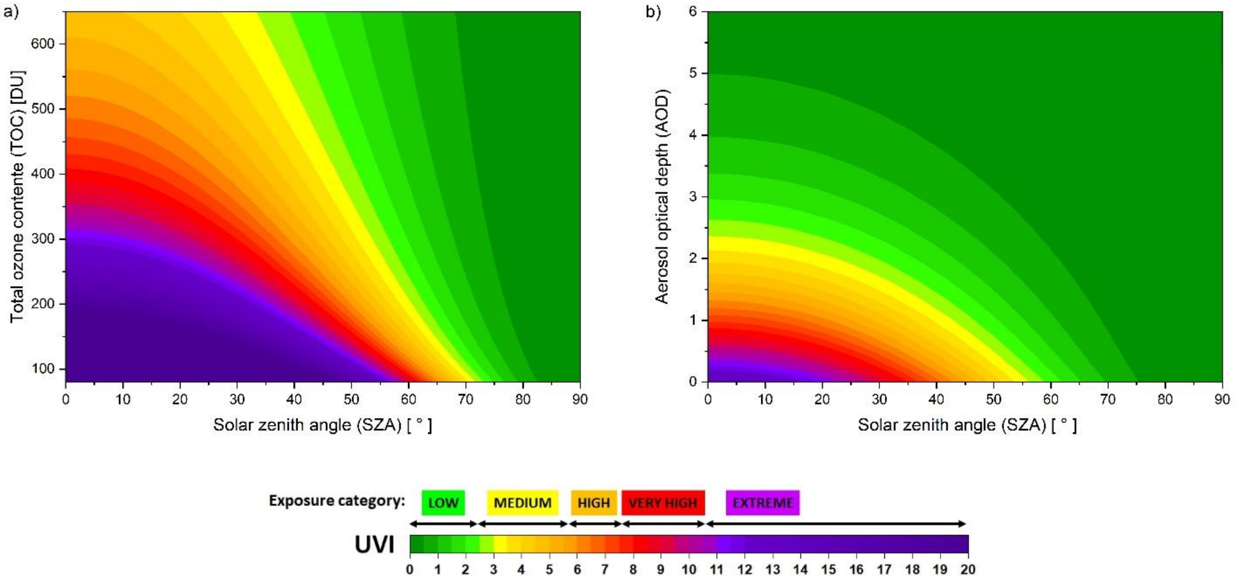

Figure 1a: UVR radiation variability according atmospheric parameters: a) UVI per variations in SZA and TOC; b) UVI per variations in SZA and AOD

2.3 Developing and applying the UVBoost model

Surface UVR irradiances were calculated for each of the 179,452 possible combinations of the input parameters (Table 1), and were separated into both random training (70%) and testing (30%) (bootstrap) groups. The most widely-used algorithms in supervised learning ML regression models, as mentioned in section 2.1, were subsequently tested. The parameters used to estimate UVR for the models were estimated using the best adjustments, which are shown in Table 2.

Table 2: ML tools, models, and parameters

Tools	Models & Parameters
sklearn.linear_model	LinearRegression()
sklearn.svm	SVR(kernel='rbf')
sklearn.tree	DecisionTreeRegressor(max_depth=5, random_state=10)
sklearn.ensemble	RandomForestRegressor(n_estimators=60, criterion='squared_error', max_depth=5, random_state=10)
xgboost	XGBRegressor (n_estimators=180, max_depth=3, learning_rate=0.05, objective = "reg:squarederror")
lightgbm	LGBMRegressor (num_leaves=50, max_depth=3, learning_rate=0.05, n_estimators=50)
catboost.core	CatBoostRegressor (iterations=1000, learning_rate=0.05, depth=3, random_state=10)

The results were analysed using k-fold cross-validation of 50 simulations from the 70/30 training and testing groups for each model, as per Table 2. The mean coefficients of determination (r^2), the root mean squared error (RMSE), and mean absolute error (MAE) were used to compare the precision among the methods. The results were evaluated using the Friedman (Sheldon et al., 1996) and Nemenyi (Demšar, 2006) statistical tests. These tests verify the differing result groups using a statistical test that rejects the null hypothesis i.e., that the data group comparisons are similar. Friedman's test is a nonparametric ANOVA equivalent for repeated measurements. It separately ranks the performance of algorithms for each dataset in order from best to worst. I also used Nemenyi's test, which is similar to Tukey's ANOVA test, to compare the classifiers with each other to complement the analyses.

Finally, two studies were run to test the accuracy and speed of the calculations to compare traditional calculation performance using RTM vs. UVBoost. In the first study, I used UVR climatological projections for the entire planet using a large database (e.g., CMIP6). This test sought to evaluate the regression models using ML as an alternative solution for calculating large databases, for high precision, and for studies requiring more time and more detailed spatial scales. In the other study, I selected eight locations at different latitudes in the northern and southern hemispheres to evaluate the accuracy of the point calculations, and to evaluate accumulated daily doses of UVR. This second analysis sought to verify the impact of accumulated errors in radiation doses obtained using irradiance integration over time intervals.

https://doi.org/10.5194/egusphere-2022-465
Preprint. Discussion started: 5 July 2022

3. Results

3.1 Validating the regression model

Table 3 shows the cross-validation statistics for the 50 simulations with different random training and testing groups, in 70% and 30% proportions, respectively.

Table 3: Cross-validation statistics (average of 30 simulations)

	RLM	SVR	AD	RF	XGB	GBM	CAT
r^2	0.416	0.994	0.893	0.914	0.979	0.991	0.998
r^2 (sd)	0.012	0.001	0.008	0.006	0.001	0.001	0.000
MAE	0.053	0.563	0.018	0.016	0.010	0.008	0.003
RMSE	0.088	0.757	0.001	0.034	0.017	0.011	0.005
rank	7.00	2.00	6.00	5.00	3.00	4.00	1.00

The Nemenyi and Friedmann tests were applied to verify the statistical significance of the proposed models. The Friedmann test showed high statistical significance (p-value << 0.001) for the differences between the coefficients of determination in the models used. Furthermore, the Critical Distance (CD) of the Nemenyi test was 1.274. The CD value indicates the minimum distance between the database ranks that is needed to constitute a significant difference between the models. In this study, the SVR and CAT models were more accurate, with no statistically significant differences between them ($CD > Rank_{(CAT)} - Rank_{(SVM)}$). However, the SVM model had MAE and RMSE greater than the CAT, and data normalization was needed, along with high computational costs for making the calculations. I chose the CAT model as the basis for UVBoost, because it showed the best results, and had one of the shortest training computational times.

3.2 Speed and accuracy testing: applying UVBoost to estimate UVR fluctuations globally

I took a set of information on atmospheric parameters, TOC, and AOD from CMIP6 (Figures 2a and 2b, respectively) to test the UVBoost model. The TOC and AOD data were extracted at a 1° x 1° spatial grid across the planet, with latitudes between ± 89.5°, and longitudes between ± 179.5°, resulting in a total of 64800 grid points. The SZA value for solar noon at each grid point was taken for the sun's position (Figure 2c). The results shown in the next figures are for the SSP370 scenario simulation (regional rivalry, with radiative forcing at 7 W/m² up to 2100 (DKRZ, 2022)), with average monthly data for March from 2021 to 2040. Similar results were obtained for other scenarios and TOC and AOD concentrations.

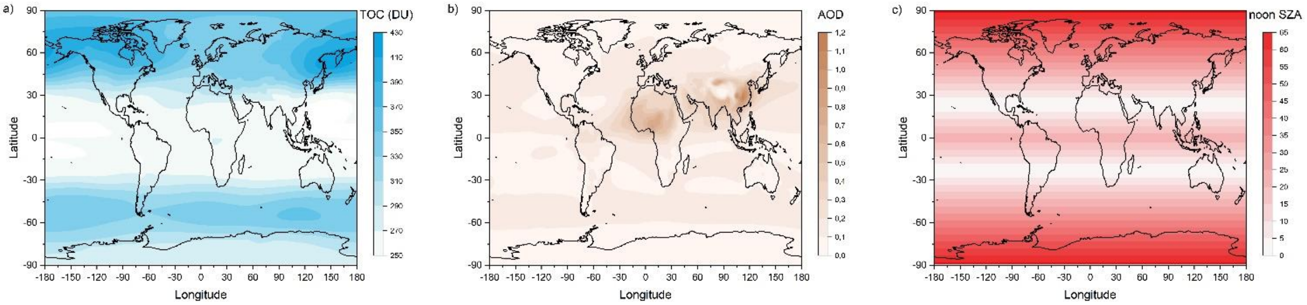

Figure 2: TOC, AOD and SZA used for the UVI calculations

The results of the UVI simulations with RTM TUV are shown in Figure 3. The results indicate extreme UVI levels at the equator and in subequatorial regions, with very high levels over Central Africa, given aerosol masses in that region. This was also observed over east-central China, and over certain areas of Southeast Asia, and in the Middle East. It is worth noting that this is an arbitrary simulation involving particular RTM input parameter conditions.

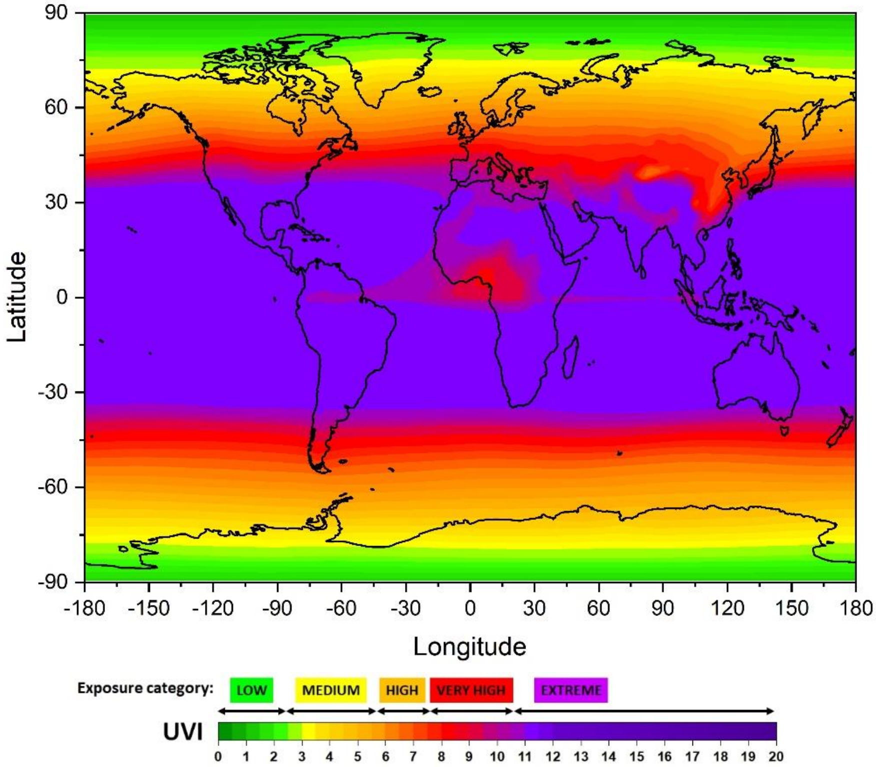

Figure 3: UVI for solar noon at each grid point (1° x 1°), using TOC and AOD values

The SZA, TOC, and AOD data from Figure 2 were used as input data in the UVBoost model. It is worth noting that the UVBoost model was first trained using the predictors as shown in Table 1. The resulting differences between the RTM TUV results and the UVBoost results are shown in Figure 4.

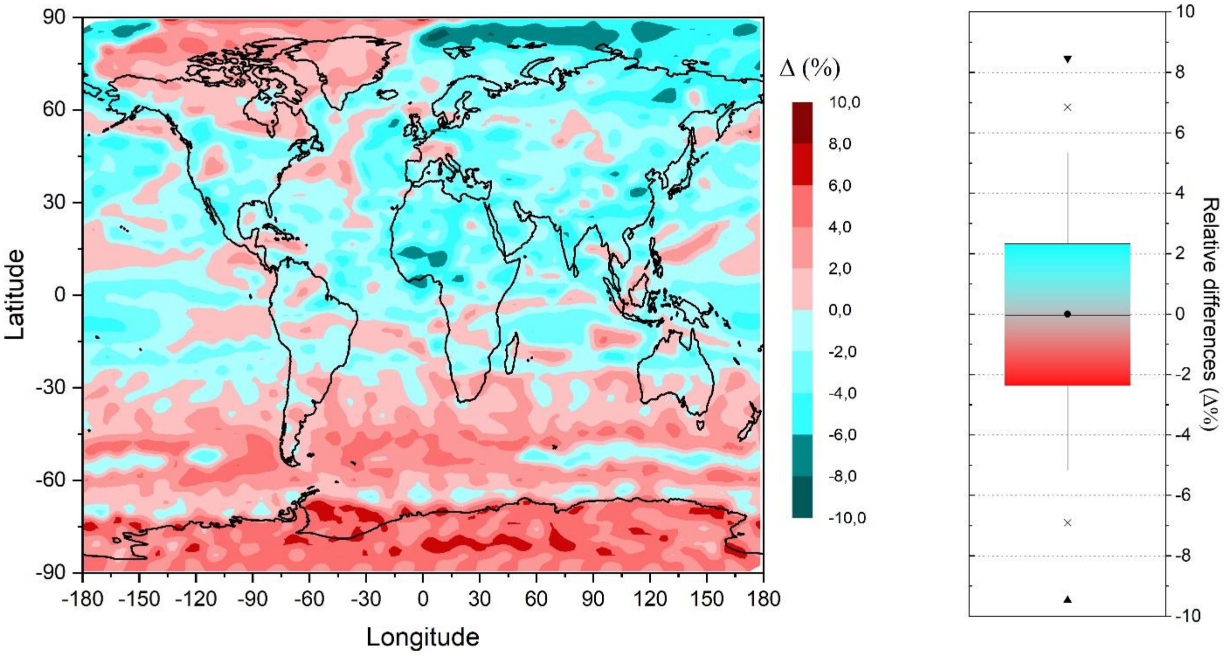

Figure 4: UVI Differences (Δ %) in RTM TUV vs. UVBoost. The boxplot on the right shows the error distribution. The whiskers limits show the 5th and 95th percentiles. The "x" indicates the 1st and 99th percentiles. The triangles show the minimum and maximum value differences.

UVBoost is a hybrid model that used a database (look-up table) taken from robust and accurate calculations obtained by RMT TUV. As was previously mentioned, the numerical regression from the SZA, TOC and AOD inputs was based on CatBoost, which is an improved regression model based on random forest techniques. The differences in the data set for the simulations were normally distributed (approximately) at a mean of 0.01%. Differences between -2.36% and +2.34% were concentrated between the 1st and 3rd quartiles of the data. The maximum errors were less than ± 10.00%, similar to traditional RTMs, and were concentrated for higher SZA, where radiation scattering processes result in more uncertainties (Lamy et al., 2019).

The small differences between the traditional RTM results and the UVBoost results corroborate the use of SML regression techniques to predict UVR fluctuations. However, the biggest and most surprising contribution of UVBoost was the significantly reduced computational times. I used a computer containing a 3.40GHz Intel(R) Core (TM) i7-6700 CPU, with 16 GB of RAM memory for the simulations. It took the RTM TUV just over 30.5 hours to calculate the high-resolution flows for the 64,800 grid points. By contrast, UVBoost took just 12.5 seconds to perform the same operation. UVBoost was over 8700

https://doi.org/10.5194/egusphere-2022-465
Preprint. Discussion started: 5 July 2022

times faster than the RTM in performing calculations with very reasonable accuracy. This is a very promising result in terms of calculation speed for large amounts of information, e.g., modeling global radiation fluctuations under different climate change scenarios in detailed time and spatial grids.

3.3 Application Test and Precision testing: Applying the UVBoost for estimating UVR doses in different locations

The second set of calculations sought to evaluate UVBoost's ability in estimating punctual and time-integrated results. UVBoost was evaluated for its predictive power in modelling UVI for a given location in this experiment, and for its predictive power in modelling integrated UVE doses, which is an important indicator of possible health risks due to excessive sun exposure. The calculations were performed at 10-minute intervals, from sunrise to sunset, at the summer solstice in the southern hemisphere (winter in the northern hemisphere) at eight different locations (Table 2). This date was chosen to allow for the greatest possible diversity of input data in the models, thereby allowing for a more comprehensive assessments of solar disk positions. I used the TOC climatological average between 2004 and 2021 as input parameters, which is available from the NASA Ozone Watch (https://ozonewatch.gsfc.nasa.gov/) system, which is collected using an Ozone Monitoring Instrument (OMI) sensor aboard the Aura satellite. I used average values from the MODIS Aerosol Product (https://modis.gsfc.nasa.gov/data/dataprod/) for AOD that are commonly observed at these locations. The simulation data are shown in Table 4.

Table 4: Location and atmospheric parameter data, TOC and AOD, for Test 2

Location	Acronym	Latitude	Longitude	TOC (DU)	AOD	noon SZA
Stockholm, Sweden	STO	59.3	18.1	300.6	0.05	82.7
New York, USA	NYK	40.7	-74.0	301.8	0.05	64.1
Shanghai, China	SHA	31.2	121.5	270.4	0.50	54.6
Abuja, Nigeria	ABU	9.1	7.5	251.2	0.60	32.5
Natal, Brazil	NAT	-5.8	-35.2	261.9	0.05	17.6
São Paulo, Brazil	SPA	-23.6	-46.6	265.7	0.05	1.0
Sydney, Australia	SYD	-33.8	151.2	284.1	0.02	10.4
Antofagasta, Chile	ANT	-62.1	-58.4	307.2	0.05	38.7

Figure 5a shows both UVI simulations, with RTM TUV (solid lines) and UVBoost (dotted lines). The daily UVI curves represent the expected Gaussian behaviour of solar irradiance estimates on a clear day. Figures 5b and 5c complemented the model comparisons. Figure 5b shows the linear regression curve for all data points from both sets. This is an almost ideal correlation of the data sets ($y = 0.99254x + 0.01566$, $r^2 = 0.9997$, p-value < 0.001). Figure 5c shows the boxplots for the relative differences between the point-to-point data of both models. The most significant differences were observed for when the solar disk was closer to the horizon (higher SZA). Since simulations were performed for winter in the northern hemisphere, SHA, NYK, and STO, had maximum solar elevations at approximately 35°, 26° and 7°, respectively, which resulted in greater differences among the models.

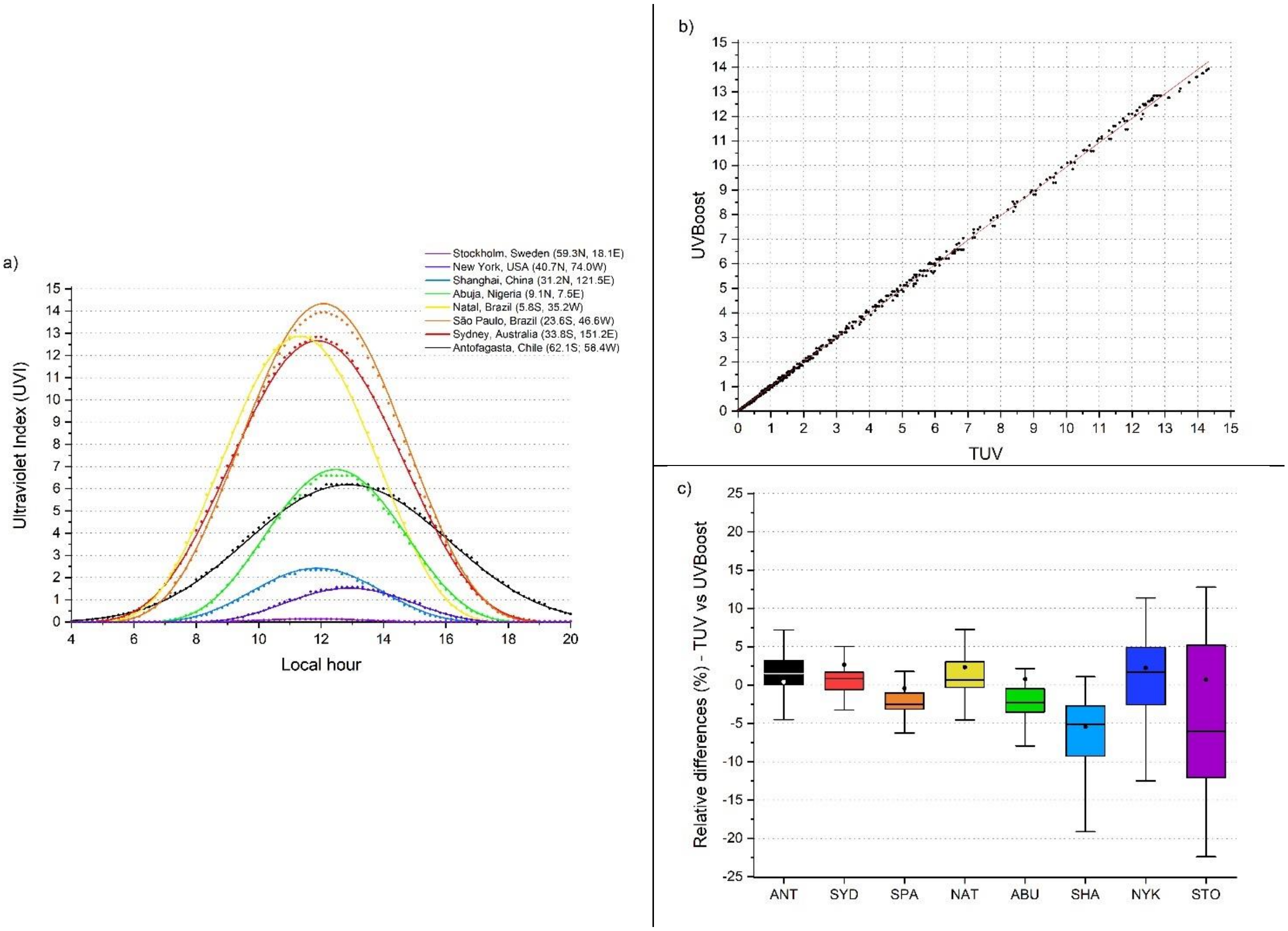

Figure 5: Analysis of the UVBoost point estimates: a) UVI estimated for December 21st at different locations on the planet: TUV (solid line), UVBoost (dotted); b) regression curve for the data sets generated by the RTM TUV and the UVBoost for the eight locations; c) boxplot of the relative differences between the data generated by the RTM TUV and the UVBoost.

The UVBoost may have had greater difficulty in estimating more tenuous UVR levels, since these situations may be associated with intense molecular scattering in the UV region. Given that it is a mathematical regression model, simulations may mistakenly associate finer UVR levels with UVR levels generally obtained for atmospheric conditions with higher TOC or AOD amounts, instead of particular situations of high SZA. In any case, lower solar radiation availability situation errors do not compromise integrated daily UVE dose assessments. Table 5 gives the integrated daily UVE doses of (Jm^{-2}) from sunrise to sunset, and ± 1h around solar noon.

Table 5: Integrated UVE doses (Jm^{-2}) throughout the day, at different locations

Daily	**ANT**	**SYD**	**SPA**	**NAT**	**ABU**	**SHA**	**NYK**	**STO**
TUV	4286	7150	7708	6453	3167	1036	628	50
CAT	4358	7225	7528	6500	3089	1003	659	72
Rel diff (%)	1.7	1.0	-2.3	0.7	-2.5	-3.1	4.9	42.8
Abs diff	71	75	-180	48	-78	-33	31	21
Noon±1h	**ANT**	**SYD**	**SPA**	**NAT**	**ABU**	**SHA**	**NYK**	**STO**
TUV	1184	2405	2715	2429	1287	451	283	27
CAT	1195	2431	2640	2436	1252	440	292	25
Rel diff (%)	1.0	1.1	-2.8	0.3	-2.7	-2.4	2.9	-8.6
Abs diff	12	26	-75	6	-35	-11	8	-2

The relative daily dose differences estimated by both models were less than 5% for most all locations, except for STO, which is located close to the arctic circle, where the total accumulated UVE doses at the winter solstice are close to zero. These are very precise estimates, especially for locations with greater solar radiation availability, where the differences in the models were less than 2.5%. There was a reduction in the STO dose differences for the integrated doses between ± 1h of solar noon. The relative differences between daily UVE doses and solar noon were similar for lower latitude locations, and for all of the southern hemisphere.

The SPA simulation is a good example of the high accuracy of the UVBoost model. At SPA, the relative errors were -2.3% for daily UVE doses, and -2.8% for noon UVE doses, i.e., the UVBoost model underestimated integrated doses by only 75 and 180 Jm^{-2}, respectively. These differences are similar to the simplified RTM model (Badosa et al., 2005), and are small, since they represent less than half of the minimum Erythematous dose needed (200 Jm^{-2}) for causing sunburns in melano-sensitive phototype I individuals, i.e., the lowest levels on the Fitzpatrick scale (Fitzpatrick, 1988). Thus, Erythematous dose predictions using the UVBoost model were quite accurate, especially under greater solar radiation availability conditions, and therefore, locations corresponding to greater concern for the sun-related health issues.

4. Conclusion and Outlook

There is consensus that RTMs estimate surface UVR with great precision, especially under clear sky conditions. However, these models are complex, and the computational time required for obtaining more accurate calculations can be high, especially when estimating over the long term for very detailed spatial grids. SML techniques for handling big data in multidisciplinary research are now popular given their practical applications, computational performance, and excellent results for regression and classification models. This study presented a hybrid model, called the UVBoost, which combined RTM precision with ML CatBoost efficiency. UVBoost consists of a robust, detailed, and accurate database, provided by an RTM, which was used to train the regression model, that allows for estimating surface UVR using information on sun position (SZA), TOC, and AOD.

Different SML tools were trained, and compared to the rest, i.e., SVR and CAT. They were statistically similar and gave superior results. However, SVR required more computational time than CAT. Furthermore, computational time grows significantly for large input data sets, since SVR only produces satisfactory results if data are normalized. Thus, SML CAT was selected as the regression model in UVBoost.

Two simulation sets were used to test UVBoost vs. traditional RTM. The first test was simulating global grids, where model speed and accuracy were tested on a large dataset. Differences between UVBoost and RTM were less than ± 5% for 90% of all calculations, and maximum differences were less than ± 10% for the entire dataset. The greatest differences were under conditions where the RTMs themselves have higher biases, i.e., high SZA, where radiation scattering is more complex. However, the most positive and surprising result was the computational speed, where UVBoost solved calculations on the 64800 grid points in a matter of seconds, meaning that this solution is about 9000 times faster than the traditional RTM.

The second test simulated UVR fluctuations from sunrise to sunset at some locations. This test sought to test the applicability and accuracy of UVBoost in determining accumulated doses. Once Again, UVBoost proved to be sufficiently accurate in making estimations for most locations, with differences less than ± 5% for most results. Higher errors were only observed at high latitude locations in the winter, when the sun's elevation during the day does not exceed 35°. Here, UVR levels were tenuous, and absolute errors were not significant enough to cause health risks. Relative errors in the order of ± 2 to 3% represented less than half of erythemal radiation doses needed for causing sunburns in the most sensitive people (phototype I) at locations closer to the equator. Thus, UVBoost proved sufficiently accurate in calculating UVR fluxes. Eventual accumulated error propagations in integrations in time proved to be negligible.

The results here constitute an important advance in contribution to disseminating UVI information to the general public, as recommended by the WHO. UVBoost and its library can enable much more accurate calculations anywhere on the planet. Since it used the Python coding language, it can be easily coupled to web pages, or with other online information tools. Furthermore, the speed of UVBoost allows climatological calculations and future scenarios forecasts to be projected with much more agility, maintaining precision and accuracy for studies of this nature. Future studies will improve upon UVBoost by expanding its database, and by including greater detailed information on atmospheric parameters, including certain information on 3D cloud models, and extended spectral band models in both the visible and infrared solar spectrum.

Code and data availability

The input CMIP6 files and the model outputs can be obtained from the author upon request. The UVBoost source code and configuration files are archived at https://doi.org/10.5281/zenodo.6783409. The SML approaches in the UVBoost development are available at https://github.com/mpcorrea-unifei/SML_for_UVBoost.

Author contributions

MPC conceived the study, carried out the research and wrote the paper.

Competing interests

The contact author has declared that neither they nor their co-author has any competing interests.

Disclaimer

Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Dedication and acknowledgments

I dedicate this work to my father, Mr. Rutênio de Paula Corrêa, who died in November 2021. I would like to thank Ph.D. student Ana Leticia Yamamoto for information on atmospheric components, and Professors Jones Granatyr, and Luciano Galdino for their valuable lessons on Python and Machine Learning. I acknowledge the World Climate Research Programme, which, through its Working Group on Coupled Modelling, coordinated and promoted CMIP6. I thank the climate modeling groups for producing and making available their model output, the Earth System Grid Federation (ESGF) for archiving the data and providing access, and the multiple funding agencies who support CMIP6 and ESGF. I would also like to thank L'Oréal for funding the SolUVCC research project, to which the UVBoost model will be specially applied, and to Vichy Laboratories for granting me the Vichy Exposome Grant 2021. Finally, I would like to thank CNPq for the scientific productivity grant (#309358/2020-5).

Financial support

The preparation of this paper was partially supported under Vichy Exposome Grant 2021 and L’Oréal France (SolUVCC research project).

References

Badosa, J., González J., Calbó J., van Weele M., McKenzie, R.L.: Using a Parameterization of a Radiative Transfer Model to Build High-Resolution Maps of Typical Clear-Sky UV Index in Catalonia, Spain, J. Appl. Meteor., 44(6), 789-803, doi: 10.1175/JAM2237.1, 2005.

Blumthaler, M.: UV Monitoring for Public Health. Int. J. Env. Res. Pub. He., 15(8), 1723, doi:10.3390/ijerph15081723, 2018.

Braiek, H.B., Khomh, F.: On testing machine learning programs. J Syst Soft, 164. doi: 10.1016/j.jss.2020.110542, 2020.

Chapman, S.: On ozone and atomic oxygen in the upper atmosphere. The London, Edinburgh, and Dublin Philos. Mag. and J. Sci., 10(64), 369-383, doi: 10.1080/14786443009461588, 1930.

Finlayson-Pitts B.J., Pitts J.N.: Chemistry of the Upper and Lower Atmosphere, Academic Press, doi: 10.1016/B978-0-12-257060-5.X5000-X, 2000.

Chen T., Guestrin, C.: XGBoost: A Scalable Tree Boosting System. KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794, doi: 10.1145/2939672.2939785, 2016.

CIE: Action spectrum for the production of previtamin D3 in human skin, International Commission on Illumination, Technical Report 174, 2006.

Demšar, J.: Statistical Comparisons of Classifiers over Multiple Data Sets. J. Mach. Learn. Res., 7(1), 1-30, https://www.jmlr.org/papers/volume7/demsar06a/demsar06a.pdf, 2006.

DeVecchi, T., Ripper, J., Roy, D., Breton, L., Marciano, A., Souza, P., Corrêa, M.P.: Using wearable devices for assessing the impacts of hair exposome in Brazil, Sci. Rep., 9, 13357, doi: 10.1038/s41598-019-49902-7, 2019.

DKRZ – Deutsches Klimarechenzentrum: The SSP Scenarios. https://www.dkrz.de/en/communication/climate-simulations/cmip6-en/the-ssp-scenarios, last access: 07 June 2022.

Emde C., Buras-Schnell, R., Kylling, A., Mayer, B., Gasteiger, J., Hamann, U., Kylling, J., Richter, B., Pause, C., Dowling, T., Bugliaro, L.: The libradtran software package for radiative transfer calculations (version 2.0.1). Geosci. Model Dev., 9(5), 1647-1672, doi: 10.5194/gmd-9-1647-2016, 2016.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geosci. Model Dev., 9, 1937-1958, doi:10.5194/gmd-9-1937-2016, 2016.

Fitzpatrick, T.B.: The validity and practicality of sun-reactive skin types I through VI, Arch. Dermatol., 124(6), 869-71, doi:10.1001/archderm.1988.01670060015008, 1988.

Foyo-Moreno I., Alados, I., Olmo, FJ., Vida, J., Alados-Arboledas, L.: On the use of a cloud modification factor for solar UV (290–385 nm) spectral range. Theor. Appl. Climatol., 68, 41-50, doi:10.1007/s007040170052, 2001.

Gorman S., Lucas, R. M., Allen-Hall, A., Fleury, N., Feelisch, M.: Ultraviolet radiation, vitamin D and the development of obesity, metabolic syndrome and type-2 diabetes, Photochem. Photobiol. Sci., 16, 362-373, doi: 10.1039/C6PP00274A, 2017.

Hancock, J.T., Khoshgoftaar, T.M.: CatBoost for big data: an interdisciplinary review. J Big Data, 7(94), doi: 10.1186/s40537-020-00369-8, 2020.

ICNIRP - The International Commission on Non-Ionizing Radiation Protection. Guidelines on limits of exposure to ultraviolet radiation of wavelengths between 180 nm and 400 nm (incoherent optical radiation). Health Phys., 87(2), 171-186. doi: 10.1097/00004032-200408000-00006, 2004.

ISO/CIE: Joint ISO/CIE Standard: Erythema Reference Action Spectrum and Standard Erythema Dose, Geneva: ISO/Vienna: CIE, ISO 17166:1999(E)/CIE S 007/E-1998, 1999.

Jégou, F., Godin-Beekman, S., Corrêa, M. P., Brogniez, C., Auriol, F., Peuch, V. H., Haeffelin, M., Pazmino, A., Saiag, P., Goutail, F., Mahé, E.: Validity of satellite measurements used for the monitoring of UV radiation risk on health, Atmos. Chem. Phys., 11, 13377–13394, doi: 10.5194/acp-11-13377-2011, 2011.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu T-Y : LightGBM: A Highly Efficient Gradient Boosting Decision Tree. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA. https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf, 2017.

Lamy, K., Portafaix, T., Josse, B., Brogniez, C., Godin-Beekmann, S., Bencherif, H., Revell, L., Akiyoshi, H., Bekki, S., Hegglin, M. I., Jöckel, P., Kirner, O., Marecal, V., Morgenstern, O., Stenke, A., Zeng, G., Abraham, N. L., Archibald, A. T., Butchart, N., Chipperfield, M. P., … Yoshida, K.: Ultraviolet radiation modelling using output from the Chemistry Climate Model Initiative, Atmos. Chem. Phys. Discuss., 19(15), 10087-10110, doi:10.5194/acp-2018-525, 2019.

Leiter, U., Keim, U., Garbe, C.: Epidemiology of Skin Cancer: Update 2019. In: Reichrath J. (eds) Sunlight, Vitamin D and Skin Cancer. Advances in Experimental Medicine and Biology, vol 1268. Springer, Cham. doi:10.1007/978-3-030-46227-7_6. 2020.

Liu H., Hu, B., Zhang, L., Zhao, X.J., Shang, K.Z., Wang, Y.S., Wang, J.: Ultraviolet radiation over China: Spatial distribution and trends, Renew. Sust. Energ. Rev., 76, 1371-1383, doi:org/10.1016/j.rser.2017.03.102, 2017.

Madronich, S., Flocke, S.: Theoretical estimation of biologically effective UV radiation at the Earth's surface, in Solar Ultraviolet Radiation - Modeling, Measurements and Effects, ed. C. Zerefos, NATO ASI Series Vol. I52, Springer-Verlag, Berlin, doi: 10.1007/978-3-662-03375-3_3, 1997.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J, Carvalhais, N, Prabhat: Deep learning and process understanding for data-driven Earth system science. Nature, 566, 195–204, doi:10.1038/s41586-019-0912-1, 2019.

Ricchiazzi, P., Yang, S., Gautier, C., Sowle, D.: SBDART: A Research and Teaching Software Tool for Plane-Parallel Radiative Transfer in the Earth's Atmosphere. Bull Am Met Soc, 79(10), 2101-2114, doi:10.1175/1520-0477(1998)079<2101:SARATS>2.0.CO;2, 1998.

Schmalwieser, A.W., Weihs, P., Schauberger, G. UV Effects on Living Organisms. In: Meyers R. (eds) Encyclopedia of Sustainability Science and Technology. Springer, New York, NY. doi: 10.1007/978-1-4939-2493-6_454-3, 2018.

Sheldon, M.R., Fillyaw, M.J., Thompson, W.D.: The use and interpretation of the Friedman test in the analysis of ordinal-scale data in repeated measures design. Physiot. Res. Int., 1(4): 221-228. doi: 10.1002/pri.66, 1996.

Sogacheva, L., Popp, T., Saye,r AM., Dubovik, O., Garay, MJ., Heckel, A., Hsu, NC., Jethva, H., Kahn, RA., Kolmonen, P., Kosmale, M., de Leeuw, G., Levy, RC., Litvinov, P., Lyapustin, A., North, P., Torres, O., Arola, A.: Merging regional and global aerosol optical depth records from major available satellite products. Atmos. Chem. Phys., 4, 2031-2056, doi:10.5194/acp-20-2031-2020, 2020.

Vuilleumier, L., Harris, T., Nenes, A., Backes, C., Vernez, D.: Developing a UV climatology for public health purposes using satellite data. Environ. Int., 146, 106177, doi: 10.1016/j.envint.2020.106177, 2021.

Wenmin, Q., Wang, L., Wei, J., Hu, B., Liang, X.: A novel efficient broadband model to derive daily surface solar Ultraviolet radiation (0.280–0.400 μm), Sci. Total Environ., 735, 139513, doi:10.1016/j.scitotenv.2020.139513, 2020.

WHO - World Health Organization, World Meteorological Organization, United Nations Environment Programme & International Commission on Non-Ionizing Radiation Protection: Global solar UV index: a practical guide, https://apps.who.int/iris/handle/10665/42459, 2002, last access: 07 June 2022.