Testing machine learning algorithms for the prediction of depositional fluxes of the radionuclides 7Be, 210Pb and 40K.

Pedro De la Torre Luque[1], Concepcion Duenas[2], Elisa Gordo[3], and Sergio Cañete[3]

[1]Istituto Nazionale di Fisica Nucleare Sezione Bari
[2]University of Malaga
[3]University of Málaga

November 23, 2022

Abstract

The monthly depositional fluxes of three natural radionuclides ($ˆ{7}$Be, $ˆ{210}$Pb and $ˆ{40}$K) were measured at a Mediterranean coastal station (Malaga, Southern Spain) over a 14-year period from 2005 to 2018, corresponding to 168 monthly samples. The study of these radionuclides provides valuable information on the atmospheric air circulation, transportation and erosion processes as well as a control of the environmental radioactivity. In this work, the depositional fluxes of these radionuclides are investigated and their relations with several atmospheric variables, such as air temperature, pressure or precipitations, have been studied by applying two popular machine learning methods: Random Forest and Neural Network algorithms. We extensively test different configurations of these algorithms and demonstrate their predictive ability for reproducing depositional fluxes of $ˆ{7}$Be, $ˆ{210}$Pb and $ˆ{40}$K. We use the Pearson-R correlation coefficient and the mean average error to evaluate the predictions of the developed models, revealing that the models derived with Neural Networks achieve slightly better results, in average, although similar, having into account the uncertainties. The mean Pearson-R coefficients, evaluated with a k-fold cross-validation method, are around 0.85 for the three radionuclides using Neural Network models, while they go down to 0.83, 0.79 and 0.8 for $ˆ{7}$Be, $ˆ{210}$Pb and $ˆ{40}$K, respectively, for the Random Forest models. Additionally, applying the Recursive Feature Elimination technique we determine the variables more correlated with the depositional fluxes of these radionuclides, which elucidates the main dependencies of their temporal variability.

Testing machine learning algorithms for the prediction of depositional fluxes of the radionuclides ^{7}Be, ^{210}Pb and ^{40}K.

P. De La Torre Luque[1], C. Dueñas[2], E. Gordo[3], S. Cañete[3]

[1]Istituto Nazionale di Fisica Nucleare, Bari, via Orabona 4, I-70126 Bari, Italy
[2]Department of Applied Physics I, Faculty of Sciences, University of Malaga, Spain
[3]Central Research Services (SCAI), University of Malaga, 29071-Malaga, Spain

Key Points:

- Machine learning methods allow us to improve predictions on environmental radioactivity.
- Correlation found between Solar cycle and depositional fluxes of crustal radionuclides ^{210}Pb and ^{40}K.
- Machine learning models adapted for studying depositional fluxes of ^{7}Be, ^{210}Pb and ^{40}K.

Corresponding author: P. De La Torre Luque, `pedro.delatorreluque@fysik.su.se`

Abstract

The monthly depositional fluxes of three natural radionuclides (^{7}Be, ^{210}Pb and ^{40}K) were measured at a Mediterranean coastal station (Malaga, Southern Spain) over a 14-year period from 2005 to 2018, corresponding to 168 monthly samples. The study of these radionuclides provides valuable information on the atmospheric air circulation, transportation and erosion processes as well as a control of the environmental radioactivity. In this work, the depositional fluxes of these radionuclides are investigated and their relations with several atmospheric variables, such as air temperature, pressure or precipitations, have been studied by applying two popular machine learning methods: Random Forest and Neural Network algorithms. We extensively test different configurations of these algorithms and demonstrate their predictive ability for reproducing depositional fluxes of ^{7}Be, ^{210}Pb and ^{40}K. We use the Pearson-R correlation coefficient and the mean average error to evaluate the predictions of the developed models, revealing that the models derived with Neural Networks achieve slightly better results, in average, although similar, having into account the uncertainties. The mean Pearson-R coefficients, evaluated with a k-fold cross-validation method, are around 0.85 for the three radionuclides using Neural Network models, while they go down to 0.83, 0.79 and 0.8 for ^{7}Be, ^{210}Pb and ^{40}K, respectively, for the Random Forest models. Additionally, applying the Recursive Feature Elimination technique we determine the variables more correlated with the depositional fluxes of these radionuclides, which elucidates the main dependencies of their temporal variability.

1 Introduction

The use of natural radionuclides as markers for studying the atmospheric circulation provides valuable information about the complex mechanisms involved. It is common to employ different natural radionuclides as tracers and chronometers in aquatic and atmospheric systems (Wogman et al., 1968; Martell, 1970; Schuler et al., 1991) and they have demonstrated to be very useful in studies dedicated to understand the mechanisms and rates of removal of aerosols (Baskaran et al., 1993). In this work, we aim at the study of a predictive model for the depositions of fallout radionuclides ^{7}Be, ^{210}Pb, and ^{40}K, whose different origins allow us to infer important features of the atmospheric circulation, erosion processes, transportation and deposition of soils and sediments from episodic to long-term timescales.

^{7}Be is a cosmogenic radionuclide originated by spallation reactions of cosmic rays with light atmospheric nuclei, such as nitrogen and oxygen (Lal et al., 1958) that has a decay half-live of $T_{1/2} = 53$ day. Thus, this nuclide is mostly produced in the stratosphere and reach the troposphere in periods of air exchange between these two layers. This is why the production of ^{7}Be is dependent on altitude, latitude and solar cycle but has negligible dependence on longitude (Baskaran et al., 1993; Dueñas et al., 2017).

In contrast, ^{210}Pb, with a decay half-live of $T_{1/2} = 22.3$ yr, is produced from the radioactive decay of ^{222}Rn, the only gaseous decay product of ^{238}U series. Therefore, ^{210}Pb is found in larger concentrations near the ground and with important dependence on the distribution of land and seas (Moore et al., 1973; Wilkening et al., 1975; Preiss et al., 1996; Garcia-Orellana et al., 2006)

The atmospheric ^{40}K ($T_{1/2} = 1.3{\cdot}10^{9}$ yr) is related to a crustal origin, from most kinds of soil, which is usually found in association with other re-suspended materials, as PM10 (particulate matter with diameter 10 μm) from the African continent (Karlsson et al., 2008; Dueñas et al., 2011).

Several works in the past have been dedicated to study the relations between the concentrations or depositional fluxes of these radionuclides with different environmental variables for different latitudes and longitudes. In this work, we employ a large dataset

Figure 1. Physical map showing the location of the study area. The zoomed window shows the exact position of study area, in Málaga.

(168 monthly measurements, from January 2005 to December 2018) of environmental variables and the fluxes of ^{7}Be, ^{210}Pb, and ^{40}K radioactivity in the Mediterranean coastal region of Málaga (Southern Spain). Similar studies were carried out in the same zone in the past and reported some important results, such as correlations with particulate material (PM10 levels) or with other environmental variables included in this work (Dueñas et al., 2004, 2009, 2011, 2017).

Here, we are exploring new methods of studying the complex relations between the depositional flux of these radionuclides and atmospheric variables, using machine learning algorithms. Machine learning (ML) techniques (Carbonell et al., 1983) provide a promising tool in the prediction of any magnitude which depends on a large number of variables and exhibits complex relations with them. Particularly, we are focused here on the implementation of these methods for the prediction of depositional fluxes of the mentioned radionuclides. These models allow us to identify subtle long-term relationships between the temporal variability of the depositional fluxes and other environmental cycles, like the Solar cycle or atmospheric cycles. Additionally, reproducing these fluxes allow us discern the real agents affecting the depositions of these radionuclides and could provide another tracer of anomalous (artificial) radiation episodes. In addition, we argue that these kind of models could be extended to different zones, always that measurements are available, to study relations with other variables not yet taken into account.

2 Materials and measurements

2.1 Study area

Málaga (4° 28' 8" W; 36° 43' 40"N), is the major coastal city in the Andalusian region situated in the south-east of Spain (see Figure 1), on the Mediterranean coast and, therefore, has a climate influenced by continental and maritime air masses. The predominant winds are easterly (SE) and westerly (NW). The climate is temperate, with contrasting wet (approximately October–April) and dry (approximately May–September) periods (Dueñas et al., 2012). The city is almost surrounded by mountains, which cause a special wind regime. Due to its geographical proximity to the African continent, our study area is frequently affected by intrusions of air masses with high concentrations of atmospheric particulate matter (Escudero et al., 2005). The sampling point is located on the flat roof of the Central Research Services (SCAI) building at the University of Málaga, at a height of 10 m above the ground and approximately at 5 km from the coastline, near the airport and surrounded by roads with traffic exhaust.

2.2 Data extraction

Bulk deposition samples were collected from January 2005 to December 2018. Samples were collected monthly using a collector that it is slightly tilted stainless steel tray 1 m^2 in area and a polyethylene vessel of 60 L capacity for rainwater sample reservoir. A volume of 6 L of the bulk deposition (the sum of wet deposition flux and the gravitational sedimentation fraction of the dry deposition) was reduced via evaporation to approximately 1 L and transferred to a Marinelli geometry container for gamma counting. The method and processing procedures were described previously (Dueñas et al., 2011). The atmospheric fluxes were calculated using the expression:

$$F = A/St \ \ (\mathrm{Bq\ m^{-2}\ month^{-1}}), \tag{1}$$

where A is the activity in the sample obtained from the gamma spectra, S is the surface area of the collector and t is the duration of sampling time. Additionally, aerosol samples were collected weekly in cellulose filters of $0.8\,\mu$m pore size and 47 mm diameter with an air sampler (Radeco, mod AVS-28A) at a flow rate of 40 l/min. A monthly composite sample containing 4 or 5 filters (depending on the number of weeks each month) was formed for the gamma analysis.

Radiometric measurements were performed by low-level gamma spectrometry with a coaxial-type germanium detector (Canberra Industries Inc., USA), with a relative efficiency of 20% and it was calibrated using certified reference gamma ray cocktail. Each sample was measured for $172,0000$ s. Gamma spectra analyses were performed with the Genie2K spectrometry software version 2.0 (Canberra Industries Inc., USA). The characteristic gamma peaks selected for the determination of the different radionuclides were: 477.6 keV for ^{7}Be, 1460.81 keV for ^{40}K and 46.5 keV for ^{210}Pb. To validate the methods, our lab routinely participates in interlaboratory comparisons to measure gamma-emitting radionuclides, in different types of samples, organized by the International Atomic Energy Agency (IAEA), the Joint Research Centre (JRC), and the Spanish Nuclear Safety Council (CSN). Further details of the low-background gamma-ray detection system have been previously described by refs. Dueñas et al. (1999, 2004).

The meteorological data (temperature, relative humidity, distance travelled monthly by the wind and precipitation) used in this study were obtained from the nearest station network of the Spanish Meteorological Agency (AEMET) (500 m away from the sampling site). Days affected by African dust outbreaks have been obtained from CALIMA project (`www.calima.es`). The monthly sunspots number were obtained from NOAA's Space Weather Prediction Center (SWPC).

Additionally, data of daily concentrations of particulate matter fraction PM10 were obtained from Carranque (36º 43' 40" N; 4º 28' 4" W), a monitoring station belonging to the regional Atmospheric Pollution Monitoring network managed by the Environmental Health Service of the Andalusian Government.

3 Methods: description of the algorithms applied and cross-validation framework

ML techniques have demonstrated their predictive power in a variety of fields, from medicine (e.g. (Lapedes et al., 1988)) to astrophysics (e.g. (Schaefer, C. et al., 2018), (Graff et al., 2014)), used for both classification (as in (Williams et al., 2006)) and numerical forecasting (see, for example, refs. Sarkar et al. (2009); vStencl and Stastny (2011)). Generally, ML methods are used to find the relation between a set of input variables and an output variable one is interested in. These variables are usually called features and labels, respectively. In the present study, the labels are the monthly depositional fluxes collected from 2005 to 2018 and the features are the atmospheric variables gathered in the same period. Earlier studies have demonstrated that it is possible to find linear relations between atmospheric variables and the depositional fluxes of these radionuclides, although the uncertainties related to this determination become too large to have accurate predictions. Using these methods we aim at obtaining more precise predictions on the depositional fluxes that could be used, e.g., to reliably detect the emission of artificial radiation or other non-expected radiation sources.

The relation between features and labels is progressively adjusted by iterating over the amount of data samples given to the algorithm, therefore the larger the amount of samples used to feed (or train) the algorithm the better the predictions become. The data sample used to adjust the algorithm is called training dataset and this adjustment process is known as the training phase, which basically consists on tuning some training parameters in order to predict the correct labels given. The algorithm adjusts itself in each iteration by comparing its predicted label with the correct label. Then, in order to evaluate the performance of the model one must provide it with new input data (i.e. these features must be different from the training data to ensure unbiased or over-fitted evaluations of the algorithm effectiveness). In this way, we can "grade" or "score" the model performance by comparing the predicted outputs with the real labels in what is called the test phase. The new set of data used in this phase is called test data.

Two different supervised algorithms have been implemented in this study; Neural Networks and Random Forest techniques, and their ability to predict depositional fluxes has been extensively tested for different configurations and for the depositional fluxes of the ^{7}Be, ^{210}Pb and ^{40}K radionuclides. Very few works have been published using ML techniques to predict depositional fluxes and none of them systematically analyzing their performance. An example of these studies can be found in ref. Chham et al. (2018), but a deeper research on the efficiency of these techniques is necessary.

The most popular ML algorithm is the *Artificial Neural Network* (ANN) model. Neural networks can learn complex patterns using layers of neurons which mathematically transform the data. The layers between the input and output are referred to as "hidden layers". A Neural Network can learn relationships between the features that other algorithms cannot easily discover, including also complex non-linear relations.

Moreover, we used an alternative and less demanding (in terms of resources) technique, the *Random Forest* algorithm [1], which, in turn, is not able to consider non-linear features in the relations between the features. This algorithm relies in an ensemble of decision trees which are combined to get averaged predictions. Each tree uses a sub-sample of the full data set, randomly selected, and progressively divides it into different nodes (or leaves) depending on certain quantitative (or qualitative, in case the tree is applied for a classification problem) criteria decided by the algorithm.

We have divided our collected data set into a training set, containing the 80-85% of the full data set, and a test set that allows us to quantify the performance of out predictions. The list of features (meteorological or atmospheric variables) employed is based on monthly averages (or monthly accumulated) and it consists of: Air temperature (in ºC), relative humidity level (%) number of days affected by African dust outbreaks (intrusions), distance travelled monthly by the wind (in km), pressure (hPa), sunspot number, amount of rainfall (dm^3), PM10 level (μg/m^3), seasonal factor (from 1, for winter, to 4, for spring), monthly factor (from 1, for January, to 12, for December), total rainfall duration (min), humid days, dry days and time between rains (in days). For both algorithms, the labels (depositional fluxes) are normalized, since this allows a better performance of the algorithm.

A Neural Network in which the input features first result into 8 units (1st hidden layer) and then into 4 units (second hidden layer) have been found to be the most adequate, as it is depicted in the Appendix A. The implementation of the Neural Network

[1] Specifically the method *RandomForestRegressor* given by the package of *sklearn.ensemble*

	^{7}Be	^{210}Pb	^{40}K
Learning rate	2.1e-3	2.1e-3	2.2e-3
Decay rate	5.e-6	5.e-5	2.4e-6

Table 1. *Main hyperparameters (i.e. the values needed to control the learning process in ML algorithms) used in the Adam optimizer, adjusted for each of the radionuclides studied.*

has been achieved by using the *Python Keras* (Chollet, 2015) library. The connections between the input features and the first hidden layer, as well as between the first and second hidden layers use the Rectified Linear Unit (ReLU) as activation function and the connections from the second hidden layer and the output units are calculated with a linear activation function.

The model performance was optimized including a step of batch normalization and dropout (finding the best results adjusting it to the 10% of the sample) after each of the hidden layers. In addition, the adaptive moment estimation optimizer, or *Adam* optimizer[2], was found to get the best performance for every one of the radionuclides. On top of this, the best results were found when taking the natural logarithm of the values for the features, as expected, and setting the mean absolute error metrics as the loss function.

Different configurations of the neural networks models and the hyperparameters involved (i.e. the values needed to control the learning process in ML algorithms) were refined by applying a simple random search method (i.e. probing different hyperparameters in an equally spaced grid of values) (Bergstra & Bengio, 2012). The optimization of the combination of these hyperparameters is left for a next work. In table 3, we show the main hyperparameters tuned for the *Adam* optimizer for each radionuclide. The rest of hyperparameters needed by the optimizer were set to their default values given by the *keras* method.

For the Random Forest algorithm, it was found that using the features values normalized, instead of their natural logarithm, gave better results. Then, the main hyperparameters were adjusted for each of the nuclides, setting the mean absolute error (MAE) as criterion for splitting the nodes and a minimum number of samples required to split an internal node (*min_samples_split*) to 3. The number of decision trees (also known as number of estimators) used in the model was set to be 680 for ^{7}Be and ^{210}Pb and 280 for ^{40}K.

The results from both algorithms and for the three radionuclides are shown and compared in the next section, in which we fully demonstrate their ability for reproducing the data and systematically explore the statistical errors around these predictions as well as the main features involved.

4 Results: predictive power of the algorithms

As a first step before running our models, we randomly shuffle the features and labels and, then, they are divided into a training and a test sets. Once the model is trained, we rate its performance by comparing the predictions with the test labels, corresponding to a 15-20% of the full data sample, using the mean percentage error and the Pearson-

[2] `https://keras.io/api/optimizers/`

R index value. While the former is an indicator of the quantitative differences between test labels and predictions, the latter is a good indicator of the trend similarities between the two sets.

In order to compare these results with a reference model, we applied the same kind of evaluation as applied for the ML algorithms to the model found in the linear regression analysis presented in ref. Dueñas et al. (2017) for the ^{7}Be radionuclide. This analysis yields a linear relation between the depositional flux of ^{7}Be and the amount of rainfall (the variable which shows the largest correlation with the depositional flux of every radionuclide) of:

$$Flux_{Be} = 6.33 + 2.6 \times \text{rainfall} \tag{2}$$

Then, the evaluation is carried out by using a portion of 25 randomly selected measurements (similar to the amount of samples in the test sets used for the ML algorithms applied) of ^{7}Be and amount of rainfall (corresponding to the same date) and measured the Pearson-R index and mean error of the predictions obtained with this reference model. In order to have a robust idea on the value of these metrics, we repeated this for 100 times (analogous to what is done in section 4.1), with different randomly selected samples of 25 measurements, and computed the average value. These metrics result in a mean R index of $\sim 0.45 \pm 0.4$ and a maximum R index of 0.95, while the mean percentage errors were of 103$\pm$150%. Having these reference metric values is necessary to compare to the quantitative results of the Random Forest and Neural Network algorithms studied here. In Figure B1 (Appendix B), we display the comparison between the predictions from the reference model and the depositional flux measurements for one of these samples.

In comparison, in Figure 2, we show some of the best results acquired from the Neural Network and Random Forest algorithms for all the studied radionuclides, which demonstrates that these algorithms can allow us to significantly improve our predictions on depositional fluxes with respect to traditional methods. Here, we highlight that these are predictions obtained from their corresponding atmospheric variables, and remark the importance of evaluating these predictions with data not used for the training phase, since this highly biases our evaluation. As we can see by the Pearson-R value, these predictions are able to suitably reproduce the labels trend with respect to the atmospheric variables. In addition, we find mean absolute errors of the order 50% usually, which are well below the error levels found using linear regressions (as shown above) and are similar to the experimental uncertainties in the determination of these fluxes, which can be $\mathcal{O}(10\%)$, as shown in refs. Herranz et al. (2008); Heydorn (2004). In this case, it has been observed that high-flux values are difficult to be matched, which may be related to periods of anomalous radiation doses. Nevertheless, this requires a dedicated study of those points and their temporal behaviour, which is beyond the scope of this paper. Further sources of uncertainty in these comparisons mainly come from the statistical uncertainties related to the measurement of the atmospheric variables and variables not included in the model.

Surprisingly, the models make good predictions also for the ^{40}K nuclide, even with a considerably smaller number of samples available for it. On top of this, we found that the absolute percentage errors follow a similar distribution for each radionuclide and both algorithms. They are well described with a Gamma probability distribution, which exhibits a slightly negative mode and a slightly positive median. This is likely due to the fact that the distribution of depositional fluxes is also be very well reproduced with a Gamma function. A representative example of these distributions for the Neural Network and Random Forest algorithms is shown in Figure 3 for the ^{7}Be radionuclide after gathering several repetitions for different test sets used. The fact that these errors follow such distribution can be used to statistically diagnose anomalous episodes of radiation doses. We noticed that the Random Forest models produce slightly larger me-

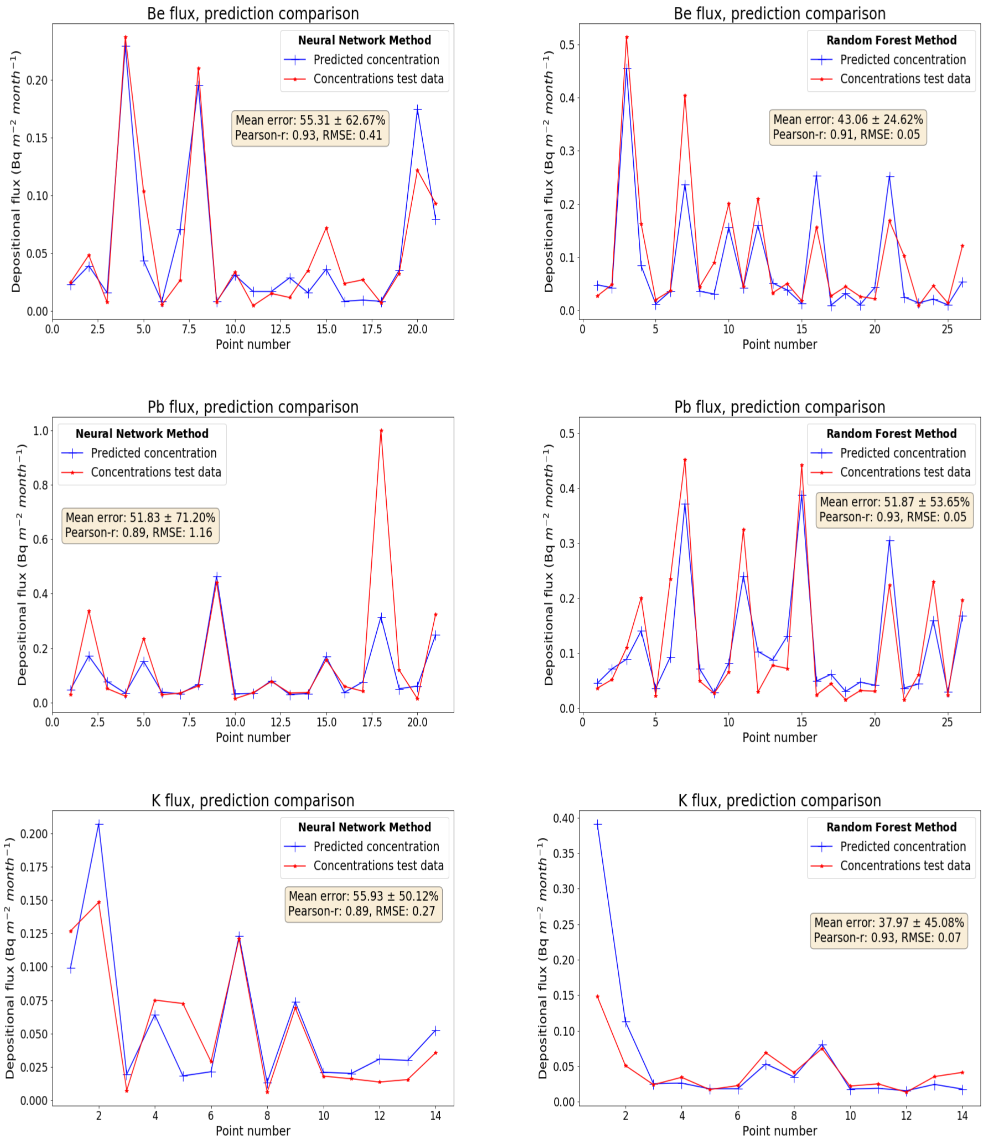

Figure 2. *Example of the results of the predictions found from the Neural Network (left panels) and Random Forest (right panels) models. These predictions are limited to the test sample, which is chosen to be around a 20% of the full data set. We also include the values of the metrics used to evaluate the predictive ability of these methods, which are the Pearson-R correlation coefficient and the mean absolute error and its standard deviation. The root mean square error (RMSE), in units of Bq m^{-2} $month^{-1}$, is also included for completeness.*

dian values and mode values more deviated from 0, but no significant differences between same algorithms for different nuclides was detected.

Nevertheless, the evaluation of the models is highly dependent on the data set used. From one side, the larger the test set, the more reliable is the model performance evaluation, but at the cost of reducing the number of samples used in the training set. On the other side, if the test set is too short, the model performance evaluation will be very uncertain. In this case, we observed that using around 20% of the full data set allowed

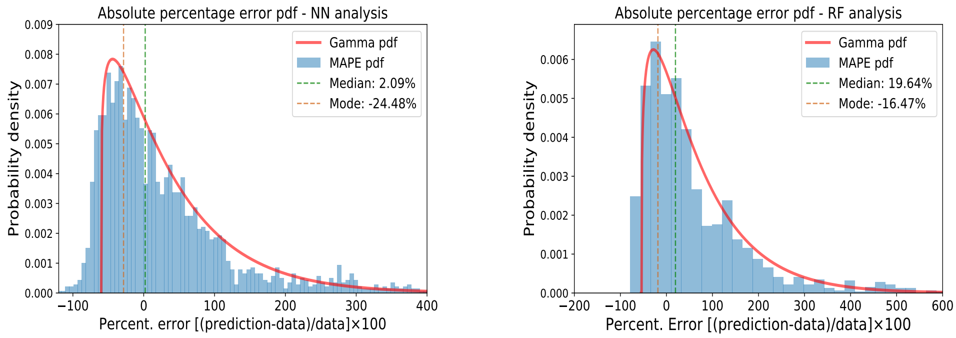

Figure 3. *Probability distribution for the percentage errors found for various evaluations with (around 20) different tests sets. The left plot shows the results of these evaluations for the Neural Network algorithm and the right plot those for the Random Forest algorithm.*

us to make consistent evaluations. Even though, they are still short enough to make our evaluation very dependent on the data test used. This issue is well known by the ML community and there are many possible strategies to deal with it to have an unbiased evaluation of our model (Raschka, 2018) and its predictions uncertainties, as it is explored in the next section.

4.1 Statistical evaluation

To prevent from biasing our model evaluation by the small amount of test data used and have into account the full uncertainty involved, we evaluate the algorithms by means of a k-fold procedure. In this process the data set is divided into k subsets. Each time, one of the k subsets is used as test set and the other k-1 subsets form the training set. Then, we statistically combine the results to get solid conclusions.

At this point, another difference between the Neural Network and the Random Forest algorithms should be taken into account to correctly manage the full uncertainties involved: while the training process exactly results in the same model for the Random Forest algorithm, this is subject to further fluctuations in the Neural Network algorithm. This is due to the optimization procedure necessary for finding the minimum error or loss when evaluating the examples in the training dataset. The main problems usually faced are: getting stacked in local minimal or local optima (i.e. regions where the loss is relatively low but it is not the lowest), saddle or flat points (regions where adjustments of the training weights do not lead to an appreciable change in the loss) and other issues more related with the loss function, gradients and the dimensionality involved. More precise information about these problems can be found, e.g., in ref. Bengio (2012). Therefore, each time the Neural Network is trained, specially when the number of samples is not large enough, it is subject to small variations in the model predictions. For this reason, a good evaluation of the uncertainties involved in the predictions of the Neural Network model requires to add these fluctuations.

In particular, we repeated the training and test phase for 5 times with the same test and training datasets. Then, we perform the evaluations with 20 different randomly-selected test sub-datasets following the k-fold procedure. This means that we carry out a total of 100 training and evaluation steps to determine the Pearson-R value and the mean percentage error of our predictions with respect to the experimental data, as well as the uncertainties related to these determinations for the Neural Network model. In turn, as the Random Forest algorithm does not suffer from those training fluctuations, we performed 60 evaluations of the model, employing a different test and training subsets, accordingly, in each evaluation.

These results are shown in Figure 4, where we represent the mean Pearson-R index values and the 1σ uncertainty related to its determination for both, the Neural Networks and Random Forest algorithms and for the three nuclides with respect to the number of iterations employed in the training phase. In general, we observe that the mean Pearson-R index values are larger for the ^{7}Be and ^{210}Pb radionuclides, while ^{40}K shows the opposite, due to the smaller number of samples available. In addition, the uncertainties related to the determination of the R index value from the Random Forest algorithm is slightly larger than that from the NN algorithm. The mean Pearson-R index values obtained are between 0.75-0.88 for ^{7}Be and ^{210}Pb, but around 0.7-0.8 for ^{40}K, although the errors are still high for every radionuclide. In particular, the determination of ^{7}Be seems to be the most accurate in general, showing a 1σ uncertainty in the determination of the R index value around ± 0.065 for the NN algorithm and ± 0.08 for the RF algorithm. A maximum mean R index value of around 0.87 and 0.88 are found for ^{7}Be and ^{210}Pb, respectively, at 1400 and 1300 iterations. The maximum mean R index value obtained for ^{40}K is slightly above 0.8, found with the RF algorithm.

As expected, the performance of these methods in reproducing depositional fluxes improves when having more samples, obtaining larger Pearson-R index values and lower uncertainties related. Nevertheless, we observed that the NN algorithm seems to accuse more the smaller number of samples with respect to the RF technique.

4.2 Selecting the main variables

To fully exploit the capability of ML techniques in improving our predictions in the depositional fluxes, we determined which are the most important features using the recursive feature elimination algorithm (RFE), which allows us to reduce the complexity and needed cpu time of the Neural Network and Random Forest algorithms and prevents from over-fitting our results. In addition, we compared the results obtained with these features with those obtained when using all the features. Specifically, we used the *RFECV* method from the $sklearn.feature_selection$ python package. The RFE algorithm is a feature selection method that allows a model to progressively eliminate the weakest features and find the best scoring combination of features.

In Figure 5 we show the optimal important features found by the RFE algorithm, along with their relative importance. As expected, the rainfall duration and rainfall volume are selected by the three radionuclides. Then, we observe that other atmospheric variables are present, as the number of humid or dry days, the average monthly pressure or the mean air temperature. On the other hand, the PM10 level and sunspot number are selected as important for the ^{40}K nuclide.

The fact that the sunspot number arises as one of the most important variables describing the depositional fluxes of ^{210}Pb and ^{40}K is unexpected. In principle, this variable is expected to be relevant for the production of ^{7}Be since it is related with the solar activity (this is, the Sun's magnetic field), which plays an important role on the flux of cosmic rays reaching the atmosphere (Yoshimori et al., 2003). This fact is probably due to the mild correlations between sunspot number and other atmospheric variables, but more data samples are needed to get a solid conclusion, since the sunspot number follows cycles of 11 and 22 years, following the solar magnetic cycles (E.W., 2015). This could be explained by the fact that there are other correlations found between the solar cycle and other atmospheric variables, as the atmospheric temperature (Qu et al., 2012) and correlations with the cosmic-ray intensity at Earth, which is known to be related to climate and involved in processes of cloud formation (Veretenenko et al., 2018; Svensmark et al., 2013; Marsh & Svensmark, 2000).

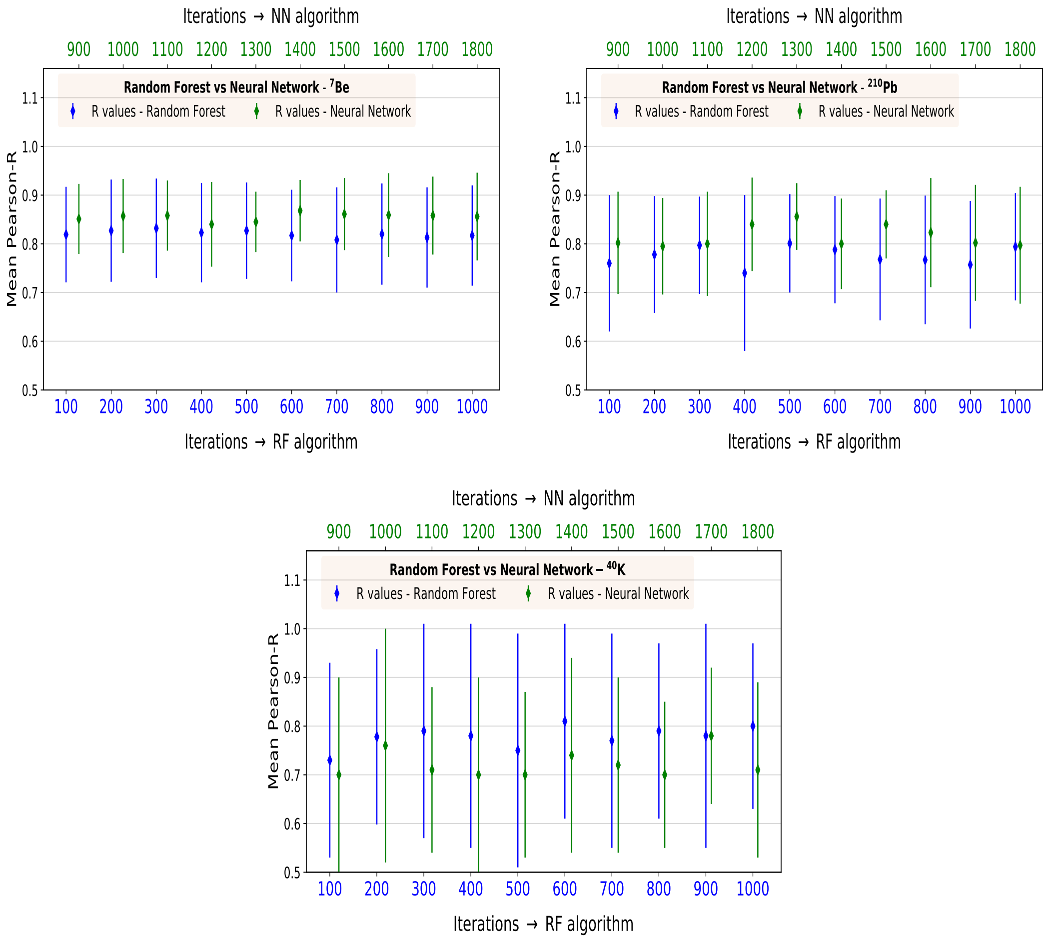

Figure 4. *Results from the k-fold evaluation of the Pearson-R correlation coefficient for the Neural Network and Random Forest algorithms for the depositional fluxes of ^7Be (upper panel), ^{210}Pb (middle panel) and ^{40}K (lower panel). The results obtained from the NN algorithm are shown in green while the results from the RF algorithm are shown in blue.*

Once these features have been selected, we proceed to compare the NN and RF algorithms explored in this work using all the features and using just the important features, as displayed in Figure 6. From this figure, we can see that the NN models for ^{40}K have significantly improved, restricting our features to be just the important ones. This means that some of the eliminated features were over-fitting the model. This can be related to the fact that this radionuclide actually comes from African zones and reach coastal zones of Southern Spain after it is transported by winds in the correct direction. Therefore, some of the atmospheric variables measured in the zone of Malaga could not be suitable to describe its amount and depositions in Malaga. Even though, the amount of rainfall should still be crucial to make the African dust to definitely fall in the study region. Furthermore, the presence of the sunspot number as an important feature have not been pointed out in the past, which may mean that there are other atmospheric variables with a considerable role in the amount and depositional flux of ^{40}K found in the Mediterranean coastal zone of the Southern Spain.

On the other hand, we see that for ^{7}Be and ^{210}Pb the results remain very similar to the case with all the features, which is quite remarkable given the number of variables

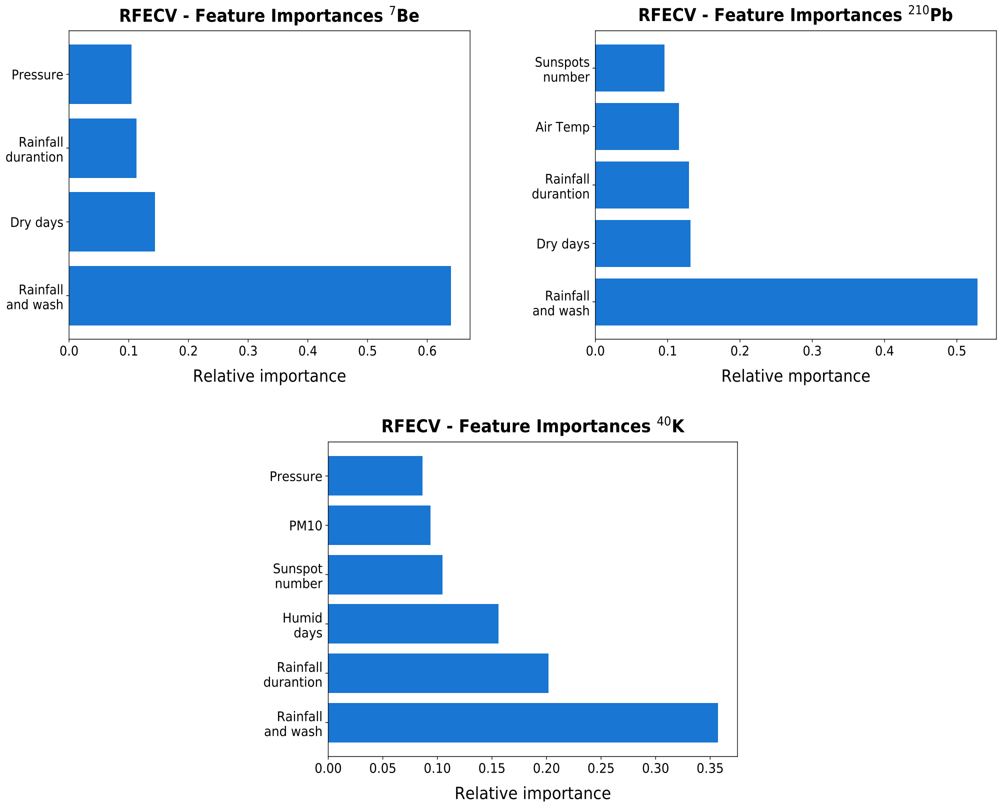

Figure 5. *Histograms with the important features found with the implemented recursive feature elimination algorithm for the depositional fluxes of ^{7}Be (upper panel), ^{210}Pb (middle panel) and ^{40}K (lower panel) with their relative importance.*

needed. In addition, the uncertainties related to the determination of the Pearson-R correlation coefficient have been considerably reduced in the NN models for ^{40}K, while they seem to be almost identical for all other cases.

In general, these results are consistent with other previously found, but the use of these ML methods allow our predictions to be more complex and better adapt to the variability related to the depositional fluxes of different radionuclides.

5 Conclusions

Modern computer algorithms allow us to refine our measurements and model predictions via new statistical tools or artificial intelligence. In this work, we have made use of two common machine learning algorithms, Neural Networks and Random Forests, in order to predict and analyse the depositional fluxes of ^{7}Be and ^{210}Pb and ^{40}K. This work has shown, first, that these methods can be successfully applied to study the depositional fluxes of different radionuclides from atmospheric variables as the amount of rainfall, pressure or air temperatures. Second, we have evaluated the performance of these models using a k-fold method and the Pearson-R coefficient and mean absolute error as metrics finding that these techniques can significantly improve old predictions made from multivariate linear regression analyses.

As expected, the performance of these methods in reproducing depositional fluxes improves when having more samples, obtaining larger Pearson-R index values and lower

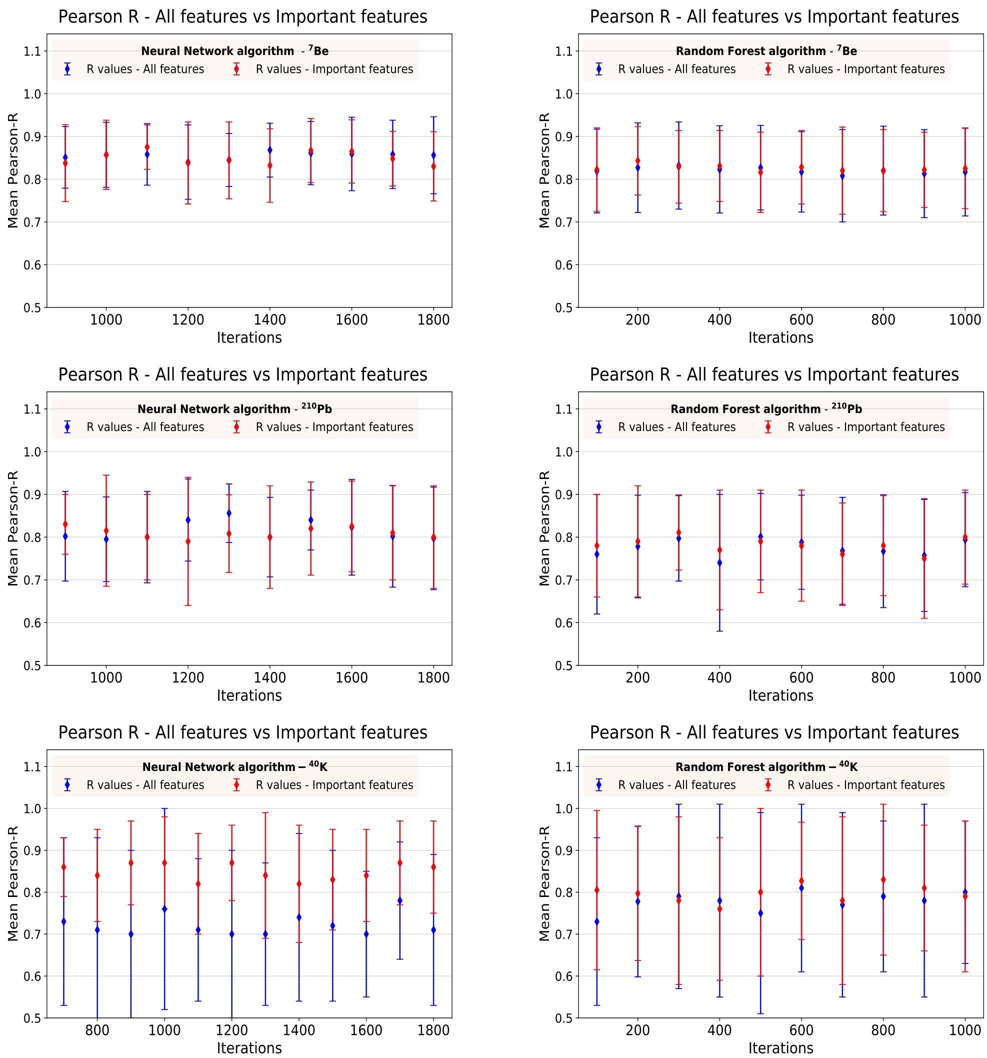

Figure 6. *Same as in Figure 4, but comparing now the results obtained using the main variables obtained from the RFE algorithm and those obtained from the models trained with all the available variables in the data set.*

uncertainties related. This, in fact, confirms the prospects on future models, with a larger number of samples measured. This is mainly related to the long times involved in the natural cycles of atmospheric variables, as, for example, the sunspot number, which is known to follow 11 or 22-years periods (solar magnetic cycles). Nonetheless, we have demonstrated that the algorithms employed here are able to reproduce the experimental depositional fluxes using monthly-averaged variables and that these predictions can help identifying periods of anomalous radiation doses. Interestingly, we found that both, the depositional fluxes of ^{210}Pb and ^{40}K, seem to be correlated with the Sunspot number.

The Neural Network models seem to reach higher mean Pearson-R index values, calculated using a k-fold cross-validation treatment, almost reaching 0.9, although the uncertainties are still quite high. Furthermore, the use of a Recursive Feature Elimination algorithm has been used to find the variables that perform the best predictions and allow us to reduce to 4, 5 and 6 the number of variables used for predicting the depositional fluxes of ^{7}Be, ^{210}Pb and ^{40}K, respectively. The training of the Neural Network and Random Forest models with these variables resulted into a negligible difference in the Pearson-R index values and the uncertainties related to its determination except for the ^{40}K nuclide in the Neural Network model, which showed a significant improvement. Even with this reduced number of variables used for training our methods, we were able to obtain mean values for the Pearson-R index value above 0.80 for all the three nuclides and both algorithms. A maximum mean R index value around 0.87 is found for ^{7}Be, ^{210}Pb and ^{40}K, respectively, at 1400, 1300 and 1200 iterations for the Neural Network method. For the Random Forest method, the maximum mean R index value of $sim0.81$ is found around 500 and 600 iterations for ^{210}Pb and ^{40}K and of almost 0.85 for the ^{7}Be radionuclide.

In conclusion, we demonstrate that Random Forest and Neural Networks methods are able to improve our current knowledge and predictions on the depositional fluxes of radionuclides in the Mediterranean coastal zone of Malaga and these models can be extended to other zones too, in order to build a more complex ensemble that could refine the existent knowledge on deposition of different radionuclides. Thus, this work constitutes the first step into the study of a large-scale (in terms of geographical areas) model able to make predictions on depositional fluxes for different geographical zones thanks to the adaptability of these algorithms. The implementation of a recurrent neural network applied to the prediction of depositional fluxes can improve these models and will be also investigated in a next work.

Acknowledgments

We would like to express our gratitude to the Consejo de Seguridad Nuclear, Spain, for their financial support to the Environmental Radioactivity Laboratory of the University of Málaga.

References

Baskaran, M., Coleman, C. H., & Santschi, P. H. (1993). Atmospheric depositional fluxes of 7be and 210pb at galveston and college station, texas. *Journal of Geophysical Research: Atmospheres*, *98*(D11), 20555-20571. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/93JD02182` doi: https://doi.org/10.1029/93JD02182

Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. In G. Montavon, G. B. Orr, & K. Müller (Eds.), *Neural networks: Tricks of the trade - second edition* (Vol. 7700, pp. 437–478). Springer. Retrieved from `https://doi.org/10.1007/978-3-642-35289-8\_26` doi: 10.1007/978-3-642-35289-8_26

Bergstra, J., & Bengio, Y. (2012, February). Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, *13*(null), 281–305.

Carbonell, J. G., Michalski, R. S., & Mitchell, T. M. (1983). Machine learning: a historical and methodological analysis. *AI Magazine*, *4*(3), 69–69.

Chham, E., Piñero-García, F., Brattich, E., El Bardouni, T., & Ferro-García, M. (2018). 7be spatial and temporal pattern in southwest of europe (spain): Evaluation of a predictive model. *Chemosphere*, *205*, 194 - 202. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0045653518307483` doi: https://doi.org/10.1016/j.chemosphere.2018.04.099

Chollet, F. (2015). *Keras.* `https://github.com/fchollet/keras`. GitHub.

Dueñas, C., Fernández, M., Carretero, J., Liger, E., & Cañete, S. (2004). Long-term variation of the concentrations of long-lived rn descendants and cosmogenic 7be and determination of the mrt of aerosols. *Atmospheric Environment*, *38*(9), 1291 - 1301. Retrieved from `http://www.sciencedirect.com/science/article/pii/S1352231003010549` doi: https://doi.org/10.1016/

j.atmosenv.2003.11.029
Dueñas, C., Fernández, M., Cañete, S., & Pérez, M. (2009). 7be to 210pb concentration ratio in ground level air in málaga (36.7^{o}n, 4.5^{o}w). *Atmospheric Research*, *92*(1), 49 - 57. Retrieved from http://www.sciencedirect.com/science/article/pii/S0169809508002342 doi: https://doi.org/10.1016/j.atmosres.2008.08.012
Dueñas, C., Fernández, M., Gordo, E., Cañete, S., & Pérez, M. (2011). Gross alpha, gross beta activities and gamma emitting radionuclides composition of rainwater samples and deposition to ground. *Atmospheric Environment*, *45*(4), 1015 - 1024. Retrieved from http://www.sciencedirect.com/science/article/pii/S1352231010009313 doi: https://doi.org/10.1016/j.atmosenv.2010.10.045
Dueñas, C., Fernández, M., Gordo, E., Cañete, S., & Pérez, M. (2012). Chemical and radioactive composition of bulk deposition in málaga (spain). *Atmospheric Environment*, *62*, 1-8. Retrieved from https://www.sciencedirect.com/science/article/pii/S1352231012007595 doi: https://doi.org/10.1016/j.atmosenv.2012.07.073
Dueñas, C., Fernández, M., Liger, E., & Carretero, J. (1999). Gross alpha, gross beta activities and 7be concentrations in surface air: analysis of their variations and prediction model. *Atmospheric Environment*, *33*(22), 3705 - 3715. Retrieved from http://www.sciencedirect.com/science/article/pii/S1352231099001727 doi: https://doi.org/10.1016/S1352-2310(99)00172-7
Dueñas, C., Gordo, E., Liger, E., Cabello, M., Ca, S., Pérez, M., & de la Torre Luque, P. (2017, 11). 7 be, 210 pb and 40 k depositions over 11 years in m alaga. *Journal of Environmental Radioactivity*, *178-179*, 325-334. doi: 10.1016/j.jenvrad.2017.09.010
Escudero, M., Castillo, S., Querol, X., Avila, A., Alarcón, M., Viana, M., ... Rodriguez, S. (2005, 09). Wet and dry african dust episodes over eastern spain. *Journal of Geophysical Research*, *110*, 18-8. doi: 10.1029/2004JD004731
E.W., C. (2015). The extended cycle of solar activity and the sun's 22-year magnetic cycle. *Space Sciences Series of ISSI*, *53*. doi: https://doi.org/10.1007/978-1-4939-2584-1_6
Garcia-Orellana, J., Sanchez-Cabeza, J. A., Masqué, P., Àvila, A., Costa, E., Loÿe-Pilot, M. D., & Bruach-Menchén, J. M. (2006). Atmospheric fluxes of 210pb to the western mediterranean sea and the saharan dust influence. *Journal of Geophysical Research: Atmospheres*, *111*(D15). Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2005JD006660 doi: https://doi.org/10.1029/2005JD006660
Graff, P., Feroz, F., Hobson, M. P., & Lasenby, A. (2014, 05). SkyNet: an efficient and robust neural network training tool for machine learning in astronomy. *Monthly Notices of the Royal Astronomical Society*, *441*(2), 1741-1759. Retrieved from https://doi.org/10.1093/mnras/stu642 doi: 10.1093/mnras/stu642
Herranz, M., Idoeta, R., & Legarda, F. (2008, 10). Evaluation of uncertainty and detection limits in radioactivity measurements. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, *595*, 526-534. doi: 10.1016/j.nima.2008.07.105
Heydorn, K. (2004, 10). Evaluation of the uncertainty of environmental measurements of radioactivity. *Journal of Radioanalytical and Nuclear Chemistry*, *262*, 249-253. doi: 10.1023/B:JRNC.0000040882.22365.a7
Karlsson, L., Hernandez, F., Rodriguez, S., Lopez-Perez, M., Hernandez-Armas, J., Alonso-Perez, S., & Cuevas, E. (2008). Using 137cs and 40k to identify natural saharan dust contributions to pm10 concentrations and air quality impairment in the canary islands. *Atmospheric Environment*, *42*(30), 7034–7042.
Lal, D., Malhotra, P. K., & Peters, B. (1958, January). On the production of ra-

dioisotopes in the atmosphere by cosmic radiation and their application to meteorology. *Journal of Atmospheric and Terrestrial Physics*, *12*(4), 306-328. doi: 10.1016/0021-9169(58)90062-X

Lapedes, A., Barnes, C., Burks, C., Farber, R., & Sirotkin, K. (1988). *Application of neural networks and other machine learning algorithms to dna sequence analysis* (Tech. Rep.). Los Alamos National Lab., NM (USA).

Marsh, N., & Svensmark, H. (2000, 11). Cosmic rays, clouds, and climate. *Space Science Reviews*, *94*, 215-230. doi: 10.1023/A:1026723423896

Martell, E. (1970). Transport patterns and residence times for atmospheric trace constituents vs. altitude. In (Vol. 93). ACS Publications. doi: 10.1021/ba-1970-0093.ch009

Moore, H. E., Poet, S. E., & Martell, E. A. (1973). 222rn, 210pb, 210bi, and 210po profiles and aerosol residence times versus altitude. *Journal of Geophysical Research (1896-1977)*, *78*(30), 7065-7075. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JC078i030p07065` doi: https://doi.org/10.1029/JC078i030p07065

Preiss, N., MéLièRes, M.-A., & Pourchet, M. (1996, December). A compilation of data on lead 210 concentration in surface air and fluxes at the air-surface and water-sediment interfaces. *Journal of Geophysical Research*, *101*(D22), 28,847-28,862. doi: 10.1029/96JD01836

Qu, W., Zhao, J., Huang, F., & Deng, S. (2012, jun). CORRELATION BETWEEN THE 22-YEAR SOLAR MAGNETIC CYCLE AND THE 22-YEAR QUASICYCLE IN THE EARTH's ATMOSPHERIC TEMPERATURE. *The Astronomical Journal*, *144*(1), 6. Retrieved from `https://doi.org/10.1088/0004-6256/144/1/6` doi: 10.1088/0004-6256/144/1/6

Raschka, S. (2018, 11). Model evaluation, model selection, and algorithm selection in machine learning. *ArXiv preprint https://arxiv.org/abs/1811.12808v3*. Retrieved from `https://arxiv.org/abs/1811.12808v3`

Sarkar, K., Ghalia, M. B., Wu, Z., & Bose, S. C. (2009). A neural network model for the numerical prediction of the diameter of electro-spun polyethylene oxide nanofibers. *Journal of Materials Processing Technology*, *209*(7), 3156 - 3165. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0924013608005827` doi: https://doi.org/10.1016/j.jmatprotec.2008.07.032

Schaefer, C., Geiger, M., Kuntzer, T., & Kneib, J.-P. (2018). Deep convolutional neural networks as strong gravitational lens detectors. *A&A*, *611*, A2. Retrieved from `https://doi.org/10.1051/0004-6361/201731201` doi: 10.1051/0004-6361/201731201

Schuler, C., Wieland, E., Santschi, P. H., Sturm, M., Lueck, A., Bollhalder, S., … Wolfli, W. (1991). A multitracer study of radionuclides in lake zurich, switzerland: 1. comparison of atmospheric and sedimentary fluxes of 7be, 10be, 210pb, 210po, and 137cs. *Journal of Geophysical Research: Oceans*, *96*(C9), 17051-17065. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/91JC01765` doi: https://doi.org/10.1029/91JC01765

Svensmark, H., Enghoff, M. B., & Pedersen, J. O. P. (2013). Response of cloud condensation nuclei (¿50 nm) to changes in ion-nucleation. *Physics Letters A*, *377*(37), 2343 - 2347. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0375960113006294` doi: https://doi.org/10.1016/j.physleta.2013.07.004

Veretenenko, S., Ogurtsov, M., Lindholm, M., & Jalkanen, R. (2018). Galactic cosmic rays and low clouds: Possible reasons for correlation reversal. In Z. Szadkowski (Ed.), *Cosmic rays* (chap. 5). Rijeka: IntechOpen. Retrieved from `https://doi.org/10.5772/intechopen.75428` doi: 10.5772/intechopen.75428

vStencl, M., & Stastny, J. (2011, 04). Artificial neural networks numerical forecasting of economic time series. In (p. 15+). IntechOpen.

Wilkening, M., Clements, W., & Stanley, D. (1975). Radon 222 flux measurements in widely separated regions. *Natural radiation environment II*, 717-730.

Williams, N., Zander, S., & Armitage, G. (2006, October). A preliminary performance comparison of five machine learning algorithms for practical ip traffic flow classification. *SIGCOMM Comput. Commun. Rev.*, *36*(5), 5–16. Retrieved from https://doi.org/10.1145/1163593.1163596 doi: 10.1145/1163593.1163596

Wogman, N. A., Thomas, C. W., Cooper, J. A., Engelmann, R. J., & Perkins, R. W. (1968). Cosmic ray-produced radionuclides as tracers of atmospheric precipitation processes. *Science*, *159*(3811), 189–192. Retrieved from https://science.sciencemag.org/content/159/3811/189 doi: 10.1126/science.159.3811.189

Yoshimori, M., Hirayama, H., Mori, S., Sasaki, K., & Sakurai, H. (2003). Be-7 nuclei produced by galactic cosmic rays and solar energetic particles in the earth's atmosphere. *Advances in Space Research*, *32*(12), 2691 - 2696. Retrieved from http://www.sciencedirect.com/science/article/pii/S0273117703800856 doi: https://doi.org/10.1016/j.asr.2003.07.006

Appendix A Sketches of Neural Network and Random Forest structures

In this appendix, we show a sketch of the general structure of the Neural Network model employed and an example of a branch of a decision tree from the Random Forest algorithm investigated in this work.

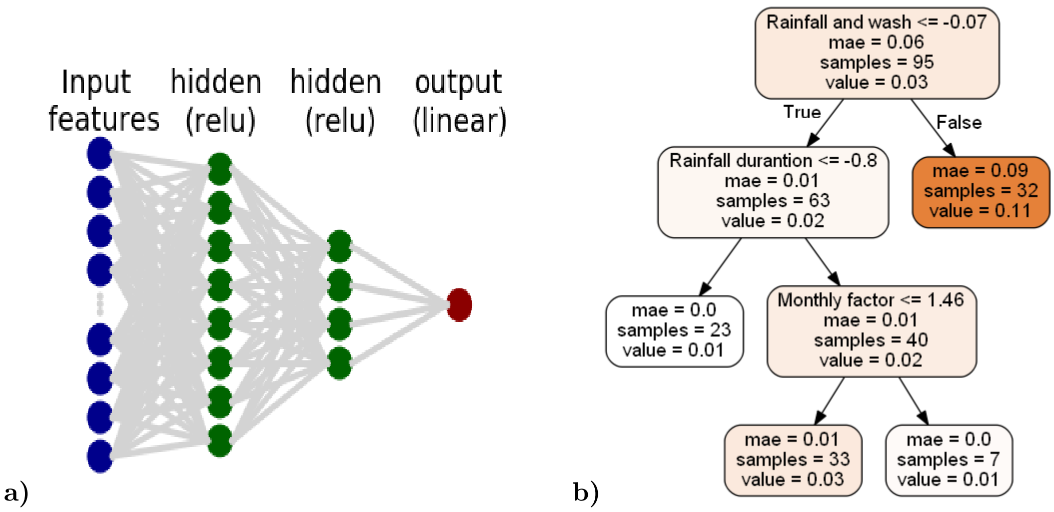

Figure A1. *a): Sketch of the Neural Network model used, where there are two hidden layers that use the ReLU activation function and an output unit that linearly combines the nodes of the last hidden layer. b): Example of a decision tree used as part of a Random Forest model.*

Appendix B Sketches of Neural Network and Random Forest structures

This appendix shows a comparison between the predictions from the reference model and the depositional flux measurements for one of these samples. It is crucial to have a reference model evaluated in the same way as for the ML algorithms studied in the paper, since this kind of evaluation is rather peculiar from ML algorithms. As we see, traditional models, based in linear regressions, are unable to reproduce the depositional fluxes behaviour, because of the complex relationships between variables.

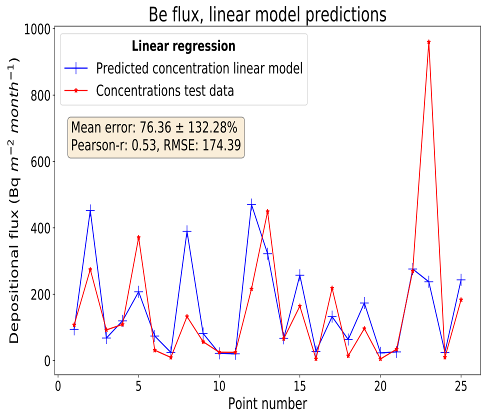

Figure B1. Predictions found from the reference linear model on one of the 25-length data samples, using the same evaluation as for the Random Forest and Neural Network algorithms studied in this work. Units of RMSE are of $Bq\ m^{-2}\ month^{-1}$.