Towards downscaling of aerosol gridded dataset for improving solar resource assessment, an application to Spain

F. Antonanzas-Torres[a,*], A. Sanz-Garcia[b], F. J. Martínez-de-Pisón[a], J. Antonanzas[a], O. Perpiñán-Lamigueiro[c,d], J. Polo[e]

[a]*EDMANS Group, Department of Mechanical Engineering, University of La Rioja, Logroño, Spain.*
[b]*Division of Biosciences. University of Helsinki, 00014 Helsinki, Finland.*
[c]*Electrical Engineering Department, EUITI-UPM, Ronda de Valencia 3, 28012 Madrid, Spain.*
[d]*Instituto de Energía Solar, Ciudad Universitaria s/n, Madrid, Spain.*
[e]*Renewable Energy Division (Energy Department), CIEMAT Avda. Complutense 22, 28040 Madrid, Spain.*

Abstract

Solar radiation estimates with clear sky models require estimations of aerosol data. The low spatial resolution of current aerosol datasets, with their remarkable drift from measured data, poses a problem in solar resource estimation. This paper proposes a new downscaling methodology by combining support vector machines for regression (SVR) and kriging with external drift, with data from the MACC reanalysis datasets and temperature and rainfall measurements from 213 meteorological stations in continental Spain.

The SVR technique was proven efficient in aerosol variable modeling. The Linke turbidity factor (TL) and the aerosol optical depth at 550nm (AOD 550) estimated with SVR generated significantly lower errors in AERONET positions than MACC reanalysis estimates. The TL was estimated with relative mean absolute error (rMAE) of 10.2% (compared with AERONET), against the MACC rMAE of 18.5%. A similar behavior was seen with AOD 550, estimated with rMAE of 8.6% (compared with AERONET), against the MACC rMAE of 65.6%.

Kriging using MACC data as external drift was found useful in generating high resolution maps (0.05ºx0.05º) of both aerosol variables. We created high resolution maps of aerosol variables in continental Spain for the year 2008.

The proposed methodology was proven to be a valuable tool to create high resolution maps of aerosol variables (TL and AOD 550). This methodology shows meaningful improvements when compared with estimated available databases and therefore, leads to more accurate solar resource estimations. This methodology could also be applied to the prediction of other atmospheric variables, whose datasets are of low resolution.

Keywords: Clear sky models, downscaling, aerosol, Linke turbidity, solar radiation, AERONET

1. Introduction

Atmospheric aerosols play a key role on the radiative energy budget of the Earth's atmosphere and contribute to the radiative forcing of the climate. Atmospheric aerosols refer to a

*Corresponding author
Email address: antonanzas.fernando@gmail.com (F. Antonanzas-Torres)

wide range of airborne particles suspended in the atmosphere with both natural and anthropogenic origin, and of quite different sizes, shapes, chemical composition and optical properties. Aerosols have a direct impact on the radiative balance due to the attenuation (scattering and absorption) of the solar radiation passing through the atmosphere. They have also an indirect contribution since aerosols can act as cloud condensation nuclei, modifying cloud properties. Knowledge of atmospheric aerosols is also of high interest for solar resource assessments in solar energy deployment, since they contribute extensively to the variability of the direct normal solar irradiance.

The aerosol loading in the atmosphere can be indirectly quantified by an optical measure named aerosol optical depth (AOD). The AERONET [1] network, which operates over 400 ground stations, provides observations of AOD with highly accurate sunphotometer sensors [2, 3]. On the other hand, satellite retrievals have been providing AOD datasets at a global scale over the last 10 years. Thus, onboard satellite instruments such as MODIS (MODerate resolution Imaging Spectroradiometer) or MISR (Multi-angle Imaging SpectroRadiometer) are able to provide AOD under cloud free conditions, as well as other particle properties, at a spatial resolution of 1x1º and 0.5x0.5º, respectively. Despite AOD's advantage of relying on observations, satellite retrievals cannot offer a complete temporal and spatial coverage of the Earth's surface as modelled estimates of AOD do. Thus, the MACC [4] (Monitoring Atmospheric Composition and Climate) project provides data records on atmospheric composition and forecast of key atmospheric constituents. The MACC aerosol reanalysis assimilates aerosol satellite retrievals to correct for model errors and it offers global and complete AOD gridded datasets at a spatial resolution of 1.125x1.125º [5, 6].

In solar resource assessments, the knowledge of AOD is of high interest since it is closely connected to the methods of estimating solar radiation components at the Earth's surface [7]. Models for estimating solar radiation components from geostationary satellite images need clear sky models, which compute solar radiation under a cloudless sky [8, 9]. Clear sky transmittance models require information on the atmospheric attenuants as input. Depending on the formulation of clear sky models, that information can be partially derived from AOD [10–12]. Some simpler models merge all the atmospheric attenuation in one unique parameter, the Linke turbidity factor, that can be estimated empirically or from the knowledge of AOD at 550 nm and the water vapor vertical column content [13–16].

Therefore, in order to better characterize the spatial and temporal variability of the aerosol loading in the atmosphere, large efforts are being taken for developing and providing complete gridded datasets at higher spatial resolutions using satellite or modelled retrievals as a reference [17, 18].

This paper focuses on increasing the spatial resolution of aerosol gridded information by using extensive ground measurements of other common meteorological variables. In particular, the work addresses the elaboration of high spatial resolution maps of AOD and the TL in Spain from the combination of MACC datasets and 213 measurement points with maximum and minimum temperatures and rainfall records using soft computing and geostatistical analysis techniques. Ground AOD data from AERONET stations located in Spain were used in both model training and final datasets assessment. The work is developed with a two-step method that first trains models to estimate aerosol variables (AOD and the Linke turbidity factor) at AERONET stations with meteorological measurements and MACC aerosol data, and then spatially generalizes these point aerosol values in the afore-mentioned 213 stations using downscaled MACC grids.

2. Data

2.1. Aerosol data

Aerosol data used in this work come from two different sources: AERONET stations located in Spain and MACC reanalysis aerosol data from 2003 to 2012. In both cases the original data are the aerosol optical depth at different wavelenghts. In case of AERONET stations, the AOD at 340, 380, 440, 500, 675, 870 and 1020 nm have been used. Likewise, the original data from MACC reanalysis were the AOD at 469, 550, 670 and 865 nm. Daily values of AOD 550 and TL have been computed from both AERONET and MACC reanalysis by using the Angstrom law and the Ineichen approach for computing Linke turbidity factor [19, 20]. In case of MACC reanalysis the final results of AOD 550 and TL have been resampled at 1x1º of spatial resolution for the geographic domain of the Iberian peninsula (37ºN to 44ºN latitude and -10ºE to 5ºE longitude).

2.2. Meteorological data

Meteorological variables are obtained from the Agroclimatic Information System for Irrigation (SIAR), a free downloadable service from which 213 meteorological stations are selected in continental Spain [21]. This network is supported by the Ministry of Agriculture, Food and Environment of Spain to derive, estimate and control different indicators for agriculture [22]. As a result, most of the stations are located in important areas for agriculture, mainly valleys and prairies and rarely in mountainous areas. Additionaly, meteorological registers in the position of AERONET stations are obtained from the Spanish Agency of Meteorology (AEMET) [23]. Figure 1 shows the locations of the stations selected from SIAR and AEMET, respectively.

Daily minimum and maximum temperatures (T_{min} and T_{max}) are recorded with *Vaisala HMP45C - Pt1000*, *Vaisala HMP155- Pt100* and *Rotronic HC2S3- Pt100* sensors with a tolerance of $\pm 0.2^{\circ}$. Relative humidity (R_h) is registered with *Vaisala HMP45C- Humicap 180*, *Vaisala HMP155- Humicap 180R* and *Rotronic HC2S3- IN-1* with $\pm 3\%$ tolerance. Rainfall (R) is measured with *ARG100* and *RM52203* sensors with a tolerance of $\pm 3\%$. The calibration of sensors is performed on a yearly basis according to the TH-007 procedure of the Spanish Centre of Metrology [24] for R_h, T_{max} and T_{min} and the IOM-84 document of the World Meteorological Organization [25] for R.

Table 1 summarizes the AERONET-AEMET database in which daily SIAR and AEMET data are collated from 2005 to 2009 and from 2005 to February 2012, respectively.

3. Methodology

The proposed methodology is described in Figure 2. In a first stage, denoted by *soft computing* techniques (Subsection 3.1), MACC and AERONET aerosol data and AEMET meteorological data are used to train *support vector machines for regression (SVR) with a radial basis kernel function* in the location of 7 AERONET stations (Table 1). The validation process of the SVR-based predictors is carried out by a 10 times repeated 10-fold cross-validation using the same AERONET measurements. In the second stage of the methodology, named *geostatistical analysis* (Subsection 3.2), MACC aerosol data is first downscaled by means of *inverse distance weighted (IDW)* interpolation and then used as the explanatory variable of *kriging with external drift (KED)* to interpolate the estimates of the SVR. This is performed for each of the 213 meteorological stations from SIAR.

3.1. Soft computing Techniques

Soft computing involves several techniques fairly tolerant with approximation, imprecision and uncertainly (artificial neural networks, fuzzy logic, probabilistic models, genetic algorithms and kernel machines, etc.) in a similar manner to mental processes in humans and is useful to solve complex not mathematically-tractable problems [26]. Kernel machines and the associated learning methods, especially the SVR, are recommended in pattern recognition due to its reliable performance showed working on many real problems [27]. In this paper, SVR have been selected to predict AOD 550 and TL. Nevertheless, this technique has proved useful in modeling different meteorological variables such as solar radiation [28] and air temperature [29].

3.1.1. Support Vector Machines for Regression (SVR)

In regression problems, given a training dataset $\{\boldsymbol{x}_i, y_i\}_{i=1}^{N}$ where $y_i \in \mathbb{R}$ and $\boldsymbol{x}_i \in \mathbb{R}^D$, [30] proposed the starting version of SVR based on Vapnik's concepts of support vector machines (Equation 1).

$$\hat{y} = \boldsymbol{w}^t \varphi(\boldsymbol{x}) + b \tag{1}$$

where $\varphi(\boldsymbol{x})$ is the mapping function from $\boldsymbol{x}$ to a higher dimensional feature space, $\hat{y}$ is an estimate of the observation y, $\boldsymbol{w} \in \mathbb{R}^D$ and $b \in \mathbb{R}$. Adjusting the model is equivalent to the problem of finding $\boldsymbol{w}$ and b that minimizes the error function subjected to some constraints (ξ_i^+, ξ_i^-). Instead of using the traditional mean square error (MSE), a more sophisticated penalty function is considered [27]. The loss function L (Equation 2), named ϵ-intensitive cost function, is zero if the predicted values are inside a ϵ-intensitive tube.

$$L = \begin{cases} 0 & \text{if } |y_i - \hat{y}_i| < \epsilon \\ |y_i - \hat{y}_i| - \epsilon & \text{otherwise} \end{cases} \tag{2}$$

The optimization of the SVR is built on inequality constraints [31], leaving a quadratic programming problem (Equations 3-6).

$$\underset{w,b,\xi,\hat{\xi}}{\text{minimize}} \quad \frac{1}{2}\boldsymbol{w}^t\boldsymbol{w} + C\sum_{i=1}^{N}\left(\xi_i + \xi_i^+\right), \;\; C > 0 \tag{3}$$

$$\text{such that} \quad w^t\varphi(\boldsymbol{x}_i) - b \leq y_i + \epsilon + \xi_i^+, \;\; i = 1, \ldots, N \tag{4}$$

$$w^t\varphi(\boldsymbol{x}_i) + b \geq y_i - \epsilon - \xi_i^-, \;\; i = 1, \ldots, N \tag{5}$$

$$\xi_i^+, \xi_i^- \geq 0, \;\; i = 1, \ldots, N \tag{6}$$

where $\boldsymbol{w}^t\boldsymbol{w}$ represents the hyperplane normal vector, C is the regularization parameter for controlling the trade-off between complexity and error and ϵ is a distance that generates the mentioned *tube* in which L is zero. In case of infinite dimensional input space, this problem cannot be solved and the issue is to calculate the optimum via Lagrange multipliers. The resulting model $\hat{y} = f(\boldsymbol{x})$ (Equation 7) will be equivalent in the feature space to Equation 1.

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b \tag{7}$$

where $K(\boldsymbol{x}_i, \boldsymbol{x}) = \varphi(\boldsymbol{x}_i)^t \varphi(\boldsymbol{x})$ is the kernel substitution applied and parameters α_i and b are determined by a linear equations system. The Gaussian kernel function (Equation 8) is selected amongst other kernel functions.

$$K(\boldsymbol{x}_i, \boldsymbol{x}) = e^{\frac{-\|x_i - x\|^2}{\sigma^2}} \tag{8}$$

where x and $\sigma \in \mathbb{R}^+$ are the center and width of the Gaussian basis function that defines the kernel behavior.

3.1.2. Feature Selection Wrapper Scheme Optimized by Genetic Algorithms

Many SVR optimizations focus on the process of feature selection (FS). For this reason, a *wrapper* scheme constituted by FS and SVR training and validation process [32] is defined to best estimate AOD 550 and TL. The selection of most significant input features and the training of the optimal SVR setting parameters are both based on genetic algorithms (GA). As a result, only the significant variables are integrated and those redundant or irrelevant are rejected as inputs for the model.

AEMET and AERONET databases are initially divided in *training-validation* and *testing* datasets in a 80/20 proportion, respectively. SVR are validated by minimizing the repeated 10 times 10-fold cross-validation relative mean absolute error ($rMAE_{cv}$). The generalization ability of these models is evaluated using the *testing* dataset to generate the relative mean absolute error of testing ($rMAE_{test}$).

The SVR are trained using GA (30 generations with a population size of 64 individuals per generation) on the *wrapper* approach, selecting the minimum number of variables that better explains the aerosol behavior. The GA select the 6 best individuals (elitism percentage of 20%) according to a fitness function, the $rMAE_{cv}$. Those individuals are also the parents for creating the next generation, using a uniform crossover operator named heuristic blending [33]. Finally, a mutation percentage of 10% is used to increase the variability of the individuals. To better understand GA and *wrapper* scheme, the authors refer to previous works [34]. Each individual is defined by a real-coded chromosome, composed by setting SVR parameters: C, σ and ϵ. The iterative procedure continues by executing the training-validation process for each SVR to adjust its weighting parameters. This process ends when the maximum number of generations has been reached or the fitness function is lower than a pre-established threshold. Eventually, the parameters of the best individual of the last generation are selected to train the final SVR with the complete dataset from AEMET and AERONET databases.

3.2. Geostatistical interpolation

3.2.1. Inverse Distance Weighted

The daily MACC rasters are previously downscaled to a higher spatial resolution (0.05x0.05°) with *inverse distance weighted* (*IDW*) interpolation [35]. The IDW has been applied in different variables mapping including metal levels in soil [36] and stratigraphics boundaries [37].

The *IDW* derives the value of the target variable at a new location as a weighted average (Equation 9), in which closer points induce a higher dependence on the sample point than further points.

$$z(s_o) = \sum_{i=1}^{n} \lambda_i(s_o) \cdot z(s_i) \tag{9}$$

where $z(s_o)$ is the target variable at the new location (o) and $\lambda_i(s_o)$ is the weight of instance i. The sum of λ must be equal to 1 in order to maintain the interpolation unbiased [38]. The weights are calculated with Equation 10.

$$\lambda_i(s_o) = \frac{\frac{1}{d^k(s_o,s_i)}}{\sum_{i=0}^{n} \frac{1}{d^k(s_o,s_i)}}; k > 1 \tag{10}$$

where k is the coefficient to adjust weights and control spatial similarity. The higher k the less dependence of further points on the estimation of the instance point. Eventually, a k of 3.5 was selected. The $d(s_o, s_i)$ denotes de Euclidean distance between o and i.

The *IDW* has two remarkable characteristics: highs and lows are limited by control points on the surface and therefore, estimates will not pass through these limits and estimates will not necessarily coincide with control points if it is not specified ,as in Equation 11 for all z_i [37].

$$z(s_o) = z_i \tag{11}$$

where z_i is any of the control points. In this study, this condition is not assumed.

This procedure enables lower the original MACC resolution of 1x1° to the final resolution of the downscaling (0.05x0.05°). As it can be seen in Figure 2, downscaled rasters with IDW are useful as explanatory variable of *kriging with external drift* in the mapping of aerosol variables.

3.2.2. Kriging with external drift

The ordinary kriging (Equation 12) is a widely used interpolation technique in which weights reflect the spatial correlation structure of control points. As a result, the estimation is defined with a global mean (μ) and a spatially correlated stochastic component ($\epsilon(s)$) by means of the semivariances between the neighboring values [39]. This technique requires three conditions which often limit its usage: a constant global mean, a constant variogram in the area studied and a target variable following a normal distribution [38].

$$z = \mu + \epsilon(s) \tag{12}$$

Kriging estimates might be improved by including information from complementary or auxiliary variables spacially sampled and correlated with the target variable, denoted as *universal kriging* or *kriging with external drift* (*KED*). Taking into consideration the correlation between MACC aerosol variables and those derivated from AERONET measurements, MACC data (previously downscaled with *IDW*) are assumed as these explanatory variables .

The *KED* includes the information from exhaustively-sampled explanatory variables in the interpolation. The estimations are performed by distinguishing between the deterministic part of variation ($\hat{m}(\mathbf{s}_\theta)$) and the stochastic spatially-autocorrelated part of variation ($\hat{\epsilon}(\mathbf{s}_\theta)$) (Equations 13 and 14).

$$\hat{z}(\mathbf{s}_\theta) = \hat{m}(\mathbf{s}_\theta) + \hat{\epsilon}(\mathbf{s}_\theta) \tag{13}$$

$$\hat{z}(\mathbf{s}_\theta) = \sum_{k=0}^{p} \hat{\beta}_k q_k(\mathbf{s}_\theta) + \sum_{i=1}^{n} \lambda_i \epsilon(\mathbf{s}_i) \tag{14}$$

where $\hat{\beta}_k$ are the coefficients estimated by the deterministic component, $q_k(\mathbf{s}_\theta)$ are the auxiliary predictors obtained from the fitted values of the explanatory variable at the new location, λ_i are the kriging weights determined by the spatial dependence structure of the residual, and $\epsilon(\mathbf{s}_i)$ are the residual at location $\mathbf{s}_i$ [22, 40].

The semivariogram is a function defined by Equation 15 ,based on a constant variance of ϵ, and also on the assumption that spatial correlation of ϵ depends on the distance amongst instances ($\mathbf{h}$) instead of their position [41].

$$\gamma(\mathbf{h}) = \frac{1}{2}\mathrm{E}(\epsilon(\mathbf{s}) - \epsilon(\mathbf{s}+\mathbf{h}))^2 \tag{15}$$

If p control points are considered, then $p \cdot (p-1)/2$ pairs of semivariances are calculated and gathered along distance to obtain the sample variogram. This sample variogram is generally fitted with a variogram model to extrapolate the spatial behavior of observed points to the area studied. In this line, different variogram functions referred such as the exponential, gaussian or spherical models are commonly refered to. In these models it is usually observed that semivariances are lower for shorter distances between pairs of control points and then stabilized to the global variance [38]. The nugget effect, generally associated with intrinsec micro-variability and measurement error, models the discontinuity of the variogram at the origin. It must be highlighted that when the nugget effect is recorded, kriging differs from a regular interpolation and as a result estimates are different from measured values [22].

3.3. Software

The methodology explained in this paper has been implemented using the open-source software R [42] and the contributed packages: `e1071` [43] for *soft-computing* analysis, `gstat` [41] and `sp` [44] for the *geostatistical* analysis, `raster` [45] for spatial data manipulation and analysis and `rasterVis` [46] for spatial data visualisation methods.

4. Results

The methodology proposed (Figure 2) is applied to downscale both AOD 550 and TL (hereinafter, aerosol variables) in continental Spain.

4.1. SVR Training

First, GA are used to seek the best SVR as well as the most relevant input variables that explain the aerosol variables. In total, 30 generations were used evaluating both MAE_{cv} and MAE_{test} in order to select the most relevant variables that explains the aerosol variables. The variable AOD 550 is explained with 5 different input variables: T_{max}, T_{min}, the daily range of temperatures of the previous day (ΔT_{i-1}), a logical variable (P) indicating if any rainfall has been registered during the day (1 for rainfall recorded and 0 for no rainfall) and AOD 550 daily data from MACC ($AOD550_{MACC}$). In the case of TL, it is explained with 6 variables: T_{min}, R, daily range of temperatures ΔT, ΔT_{i-1}, the logical variable associated to rainfall of the previous day (P_{i-1}) and the TL derived from MACC (TL_{MACC}). Temperatures and rainfall variables are included because of their inter-relationships with aerosol variables. Therefore, a higher daily range of temperatures and rainfall in the previous day is related with a cleaner and cloud-free atmosphere [47].

Table 2 shows the models' parameters from the best individual of generation 30 of both SVR trained. These parameters correspond to the parameters in the final models. The results of the *SVR training* and *validation* stages are shown in Figures 3 and 4. In both aerosol variables, similar MAE_{cv} and MAE_{test} are obtained in the last generation. However, a better modeling is achieved for AOD 550 than for TL with relative mean absolute error of testing ($rMAE_{test}$) of 8.49% and 10.57%, respectively. It can be deduced that within the first 10 generations the rMAE is stabilized and only a very slight improvement is obtained in next generations. The variation of number of features seen in TL (Figure 4) might be explained by the seeking algorithm of the *wrapper scheme* looking for the global minimum.

Tables 3 and 4 show different rMAE and coefficients of determination (R^2) evaluated in AERONET stations. The rMAE and R^2 between AERONET and MACC databases ($rMAE_{aer,macc}$ and $R^2_{aer,macc}$) show significant differences between these databases with both aerosol variables, which were also detected in other positions [48]. The average $rMAE_{aer,macc}$ and $R^2_{aer,macc}$ for TL

and AOD 550 are 65.6% and 18.5% and 0.343 and 0.069, respectively. The latter differences between estimated databases and measured data motivate the downscaling proposed integrating on-ground measured data. Scatter plots (Figures 5 and 6) of both aerosol variables are shown between AERONET and MACC and SVR. These plots also show a better performance in TL estimated with SVR than for AOD 550.

The rMAE and R^2 obtained between the SVR estimates and AERONET ($rMAE_{down,aer}$ and $R2_{down,aer}$) are significantly improved compared with $rMAE_{aer,macc}$ and $R2_{aer,macc}$. The average $rMAE_{down,aer}$ and $R2_{down,aer}$ for TL and AOD 550 are 10.2% and 8.6% and 0.8 and 0.551, respectively. The high generalization capacity achieved by SVR-trained is remarkable, especially given the various climates and low range of errors at the stations we evaluated. MACC data is considered an input feature in SVR models due to the resulting improvement in rMAE and R^2 between the SVR and MACC ($rMAE_{down,macc}$ and $R2_{down,macc}$). As a result, SVR models moderate differences between estimated and measured data.

The goodness of the *KED* is evaluated calculating the cross-validation rMAE and R^2 ($rMAE_{cv,ked}$ and $R^2_{cv,ked}$), which implies leaving the station analyzed out of the *KED* and comparing results obtained against SVR estimated data. It is noteworthy that Station 6, which corresponds to Barcelona and it is the further station, presents the highest value of $rMAE_{cv,ked}$ (Figure 1). This fact might be expected from the semi-variogram effect of kriging with distance (distant points are usually more difficult to estimate given the growing trend of the semi-variograms).

4.2. SVR Estimation and KED

The SVR models are then used to estimate daily aerosol variables in 213 meteorological stations of SIAR using their registers of meteorological variables and MACC data. Semi-variograms of both aerosol variables are adjusted with a pure nugget model, indicating a strong intrinsic variability independent from the distance between points pairs. Figure 7 shows the $rMAE_{cv,ked}$ and $rMAE_{down,macc}$ for both aerosol variables. It is remarkable that 58.2% (TL) and 88.7% (AOD 550) of the stations fulfill with lower rMAE than the average $rMAE_{down,macc}$ (Tables 3 and 4) evaluated at AERONET stations). The $rMAE_{cv,ked}$, 56.3% (TL) and 68.7% (AOD 550) of the stations present lower rMAE than the average $rMAE_{cv,ked}$ at AERONET stations. While the range of $rMAE_{down,macc}$ for TL is enclosed up to 25.2%, the range for AOD 550 rises to 70.2%.

The comparison between the downscaling proposed and original MACC data is also shown with Hovmöller plots [49]. These plots collate spatial-temporal data into a single graph averaging values latitudinally or longitudinally. Figure 8 shows Hovmöller plots of daily relative differences of aerosol variables for the year 2008 on a latitudinal basis. The TL variable shows much lower relative difference extremes (-44.3% to 51.0%) than the AOD 550 (-98% to 157.8%). It can also be deduced from Figure 8 that AOD 550 shows higher latitude dependence than TL, in which relative differences show similar values regardless of the latitude. A deep temporal influence is recorded in both aerosol variables for the range of latitudes evaluated.

Figure 9 shows the intra-pixel annual relative difference between the downscaled aerosol variables and MACC in 2008. With TL downscaling, annual relative differences are lower than 8% for the majority of Spain in latitudes below 41º. The effect of the centralization of the *IDW* is perceptible in some pixels (for instance, 39.5ºN, 5.5ºW), since the gridded MACC values are assigned to the geometrical center of each raster cell by the authors. However, the effect of the estimated values on the *KED* prevails over the *IDW* in most of the pixels of Northern Spain (latitudes over 42º). The AOD 550 mapping generates the opposite tendency to TL: higher differences are recorded for latitudes below 42º, which in some areas are recorded up to 70%. According to the average $rMAE_{aer,macc}$ and $rMAE_{down,macc}$ for AOD 550 (Table 4), 65.6% and 36.8%, respectively, a similar behavior is obtained (Figure 9). It can also be deduced from this figure that intrapixel differences are negligible in Southern Spain. Annual relative differences

present a bias, TL downscaled (TL_{down}) is lower (negative differences) than TL from MACC (TL_{macc}) and AOD 550 downscaled ($AOD550_{down}$) is higher (positive differences) than AOD 550 from MACC ($AOD550_{macc}$).

Finally, we generated maps of mean values of aerosol variables for 2008. Lower values of TL and AOD 550 are consequently obtained in Central Spain, where the atmosphere is typically clear, and higher aerosol variables are obtained in Northern Spain and the Mediterranean Basin, due to the higher water content in the atmosphere.

5. Conclusions

The elaboration of high spatial resolution maps of TL and AOD 550 in Spain is addressed combining *SVR* and *KED* with data from MACC reanalysis datasets and 213 meteorological stations with measurements of temperatures and rainfall.

The large differences between measured aerosol variables from AERONET and estimated from MACC database motivated the downscaling of MACC by combining data from both databases. The *SVR* methodology has proved efficient in aerosol variables modeling, generating significantly lower errors in AERONET positions than MACC reanalysis estimates when compared with AERONET measurements. The TL is estimated with $rMAE_{down,aer}$ of 10.2% versus the $rMAE_{aer,macc}$ of 18.5%. A similar behavior is obtained with AOD 550, which is estimated with $rMAE_{down,aer}$ of 8.6% versus the $rMAE_{aer,macc}$ of 65.6%.

The capacity of generalization of the *SVR* trained is also remarkable, given that 58.2% and 88.7% of the 213 stations (TL and AOD 550, respectively) present lower $rMAE$ than the average $rMAD_{down,macc}$ at AERONET positions. As a result, it is possible to estimate the aerosol variables with a certain level of certainty if registers of temperatures and rainfall are available.

The geostatistical analysis with *IDW* and *KED* is useful to generate high resolution maps (0.05ºx0.05º) of aerosol variables. A significantly higher heterogeneity on daily differences is obtained with AOD 550 (-98% to 157.8%) than with TL (-44.3% to 51.0%). The trend in differences between the downscaling proposed and MACC also remains lower than $rMAE_{down,macc}$ for TL with regard to annual relative differences (<78.0% with AOD 550 and <13.6% with TL).

The methodology proposed proved useful in elaborating high resolution maps of aerosol variables, which are essential in generating reliable solar radiation estimations with clear sky models. Due to the high influence of input variables on models trained, we recommend using the methodology proposed to generate new locally trained SVR models with local selection of input variables to estimate aerosol content in the region of study. Furthermore, this methodology could be also interesting to predict other atmospheric variables, whose datasets are provided with low resolution.

Acknowledgements

Fernando Antonanzas-Torres has a fellowship FPI at University of La Rioja (Spain). We are indebted to the Instituto de Estudios Riojanos for funding parts of this research. A. Sanz-Garcia would also like to acknowledge the research founding No. 273689 (FINSKIN) from the Academy of Finland.

References

[1] Aerosol Robotic Network (AERONET)
URL http://aeronet.gsfc.nasa.gov

[2] Dubovik, O., Holben, B., Eck, T. F., Smirnov, A., Kaufman, Y. J., King, M. D., Tanre, D. and Slutsker, I., 2002. Variability of absorption and optical properties of key aerosol types observed in worldwide locations. J. Atmos. Sci. 59, 590-608.

[3] Holben, B. N., Eck, T. F., Slutsker, I., Tanré, D., Buis, J. P., Setzer, A., Vermote, E., Reagan, J. A., Nakajima, T., Kaufman, Y. J., Jankowiak, I. and Smirnov, A., 1998. AERONET - A federated instrument network and data archive for aerosol characterization. Remote Sens. Environ. 66, 1-16.

[4] Monitoring Atmospheric Composition and Climate (MACC)
URL http://www.copernicus-atmosphere.eu

[5] Benedetti, A., Morcrette, J. J., Boucher, O., Dethof, A., Engelen, R. J., Fisher, M., Flentje, H., Huneeus, N., Jones, L., Kaiser, J. W., Kinne, S., Mangold, A., Razinger, M., Simmons, A. J. and Suttie, M., 2009. Aerosol analysis and forecast in the European Centre for Medium-Range Weather Forecasts Integrated Forecast System: 2. Data assimilation. J. Geophys. Res. 114, D13205.

[6] Morcrette, J. J., Boucher, O., Jones, L., Salmond, D., Bechtold, P., Beljaars, A., Benedetti, A., Bonet, A., Kaiser, J. W., Razinger, M., Schulz, M., Serrar, S., Simmons, A. J., Sofiev, M., Suttie, M., Tompkins, A. M. and Untch, A., 2009. Aerosol analysis and forecast in the European Centre for Medium-Range Weather Forecasts Integrated Forecast System: Forward modeling. J. Geophys. Res. 114, D06206.

[7] Stoffel, T., Renne, D., Myers, D., Wilcox, S., Sengupta, M., George, R. and Turchi, C., (Report 2010). Concentrating Solar Power. Best Practices Handbook for the Collection and Use of Solar Resource Data. NREL/TP-550-47465, National renewable Energy Laboratory, Golden CO.

[8] Polo J., Zarzalejo, L. F. and Ramirez, L. (2008). Solar radiation derived from satellite images, Chap. 18. In: Modeling Solar Radiation at the Earth Surface. Edited by: Viorel Badescu. Springer-Verlag.

[9] Renne, D., 1999. Overview of techniques for estimating surface solar radiation from satellites. Proceedings of: 2nd Workshop on satellites for solar energy assessments, Golden (USA).

[10] Gueymard, C. A., 2003. Direct solar transmittance and irradiance predictions with broadband models. Part I: detailed theoretical performance assessment. Sol. Energy. 74, 355-379.

[11] Gueymard, C. A., 2003. Direct solar transmittance and irradiance predictions with broadband models. Part II: validation with high-quality measurements. Sol. Energy. 74, 381-395.

[12] Ineichen, P., 2006. Comparison of eight clear sky broadband models against 16 independent data banks. Sol. Energy. 80, 468-478.

[13] Linke, F., 1922. Transmissions-Koeffizient und Trübungsfaktor. Beitr. Phys. fr. Atmos. 10, 91-103.

[14] Remund, J., Albuisson, M., Lefèvre, M. and Wald, L., 2002. Monthly mean linke turbidity world maps. Sophia Antipolis (France), Groupe Teledetection & Modelisation - Centre d'Energetique. Ecole des Mines de Paris.

[15] Remund, J. and Domeisen, D., (Report 2010). Aerosol optical depth and Linke turbidity climatology. Meteotest Report for the IEA SHC Task 36.

[16] Rigollier, C., Bauer, O. and Wald, L., 2000. On the clear sky model of the ESRA – European Solar Radiation Atlas – with respect to the heliosat method. Sol. Energy. 68, 33-48.

[17] Gueymard, C. A. and Thevenard, D., 2009. Monthly average clear-sky broadband irradiance database for worldwide solar heat gain and building cooling load calculations. Sol. Energy. 83, 1998-2018.

[18] Ruiz-Arias, J. A., Dudhia, J., Gueymard, C. A. and Pozo-Vázquez, D., 2013. Assessment of the Level-3 MODIS daily aerosol optical depth in the context of surface solar radiation and numerical weather modeling. Atmos. Chem. Phys. 13, 675-692.

[19] Iqbal, M., 1983. An introduction to solar radiation. Academic Press Canada.

[20] Ineichen, P., 2008. Conversion function between the Linke turbidity and the atmospheric water vapor and aerosol content. Sol. Energy. 82, 1095-1097.

[21] Servicio de Información Agroclimática del Regadío (SIAR)
URL www.marm.es/siar

[22] Antonanzas-Torres, F., Cañizares, F., Perpiñán, O., 2013. Comparative assessment of global irradiation from a satellite estimate model (CM SAF) and on-ground measurements (SIAR): a Spanish case study. Renew. Sust. Energ. Rev. 21 248-261.

[23] Agencia Estatal de Meteorología (AEMET)
URL www.aemet.es

[24] Spanish Centre of Metrology (CEM)
URL http://www.cem.es

[25] World Meteorological Organization (WMO)
URL http://www.wmo.int

[26] Zadeh, L. A., 1994. Fuzzy logic, neural networks, and soft computing. Commun ACM 37 (3), 77-84.

[27] Vapnik, V. N., 1995. The nature of statistical learning theory. Springer-Verlag New York, Inc., New York, NY, USA.

[28] Chen, J. L., Liu, H. B, Wu, W., Xie, D. T., 2011. Estimation of monthly solar radiation from measured temperatures using support vector machines - A case study. Renew. Energ. 36, 413-420.

[29] Paniagua-Tineo, A., Salcedo-Sanz, S., Casanova-Mateo, C., Ortiz-García, E.G., Cony, M.A., Hernández-Martín, E., 2011. Prediction of daily maximum temperature using a support vector regression algorithm. Renew. Energ. 36, 3054-3060.

[30] Drucker, H., Burges, C. J., Kaufman, L., C., C.J., Kaufman, B. L., Smola, A., Vapnik, V., 1996. Support Vector Regression Machines.

[31] Smola, A. J., Schölkopf, B., 2004. A tutorial on support vector regression. Stat. Comput. 14 (3), 199-222.

[32] Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res., 3, 1157-1182.

[33] Michalewicz, Z., Janikow, C.Z., Handling constraints in genetic algorithms. In: ICGA. pp. (1991) 151–157.

[34] Sanz-Garcia, A., Fernandez-Ceniceros, J., Antonanzas-Torres, F., Martinez-de-Pison-Ascacibar, F. J., 2014. Parsimonious Support Vector Machines Modelling for Set Points in Industrial Processes Based on Genetic Algorithm Optimization. In Á. Herrero, B. Baruque, F. Klett, A. Abraham, V. Snášel, A. C. P. L. F. Carvalho, P. G. Bringas, I. Zelinka, H. Quintián, E. Corchado (Eds.), International Joint Conference SOCO'13-CISIS'13-ICEUTE'13 (Vol. 239, pp. 1-10): Springer International Publishing.

[35] Shepard, D. A two-dimensional interpolation function for irregularly-spaced data, In: Blue, R. B. S., Rosenberg, A. M. (Eds.), Proceedings of the 1968 ACM National Conference. ACM Press, New York, pp. 517–524

[36] Bech, J., Tume, P., Sánchez, P., Reverter, F., Bech, J., Lansac, A., Longan, L., Oliver, T., 2011. Levels and pedogeochemical mapping of lead and chromium in soils of Barcelona province (NE Spain). J. Geochem. Explor. 109, 104-112.

[37] Bartier, P., Keller, C. P., 1996. Multivariate interpolation to incorporate thematic surface data using inverse distance weighting (IDW). Comput. Geosci. 22 (7), 195-799.

[38] Hengl, T. A practical guide to geostatistical mapping, 2009.
URL http://spatial-analyst.net/book/

[39] Matheron, G., 1962. Traité de géostatistique appliquée. Vol. 14 of Mémoires du Bureau de Recherches Géologiques et Minières. Editions Technip, Paris.

[40] Webster, R., Oliver, M. A., 2001. Geostatistics for Environmental Scientists. Statistics in Practice. Wiley, Chichester, p. 265.

[41] Pebesma, E. J., 2004. Multivariable geostatistics in S: the gstat package, Comput. Geosci. 30, 683-691.

[42] R Development Core Team, 2012. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
URL http://www.R-project.org

[43] Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., 2012. e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.6-1.
URL http://cran.r-project.org/package=e1071

[44] Pebesma, E. J., Bivand, R. S., November 2005. Classes and methods for spatial data in R. R News 5 (2), 9–13.
URL http://cran.r-project.org/doc/Rnews/

[45] Hijmans, R. J., van Etten, J., 2012. raster: Geographic analysis and modeling with raster data.
URL http://cran.r-project.org/web/packages/raster/

[46] Perpiñán, O., Hijmans, R., 2012. rasterVis: Visualization methods for the raster package. R package version 0.10-9.
URL http://cran.r-project.org/package=rasterVis

[47] Antonanzas-Torres, F., Sanz-Garcia, A., Martínez-de-Pisón-Ascacíbar, F. J., Perpiñán-Lamigueiro, O., 2013. Evaluation and improvement of empirical models of global solar irradiation: case study northern Spain. Renew. Energ. 60, 604-614.

[48] Polo, J., Antonanzas-Torres, F., Vindel, J. M., Ramirez, L., 2014. Sensitivity of satellite-based methods for deriving solar radiation to different external atmospheric input. In press.

[49] E. Hovmöller, The trough-and-ridge diagram, Tellus 1 (2) (1949) 62–66.

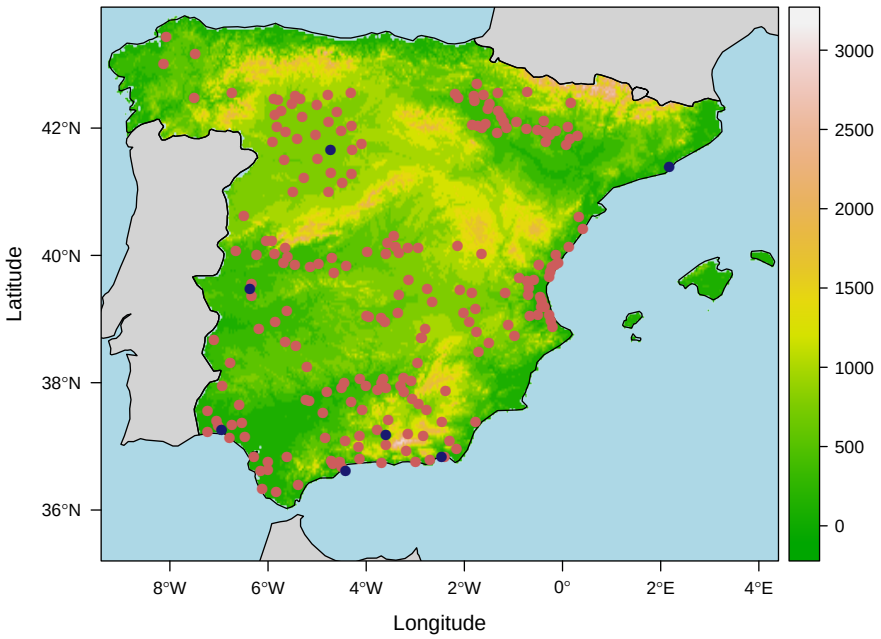

Figure 1: Topographical map of continental Spain with the 213 meteorological stations selected from SIAR (red points) and 7 stations of AEMET (blue points), whose coordinates coincide with AERONET stations.

Name	Lat.	Lon.	Elev.	Period	Instances
Caceres	39.47	-6.37	448	25/07/05 - 29/02/12	1345
Huelva	37.26	-6.94	38	17/03/10 - 29/02/12	591
Malaga	36.62	-4.42	29	06/11/08 - 29/02/12	1037
Granada	37.18	-3.60	702	01/01/05 - 29/02/12	1705
Almeria	36.84	-2.46	42	18/02/11 - 29/02/12	338
Barcelona	41.39	2.17	20	01/01/05 - 29/02/12	1939
Valladolid	41.65	-4.72	702	05/06/08 - 29/02/12	100

Table 1: Summary of AERONET-AEMET database

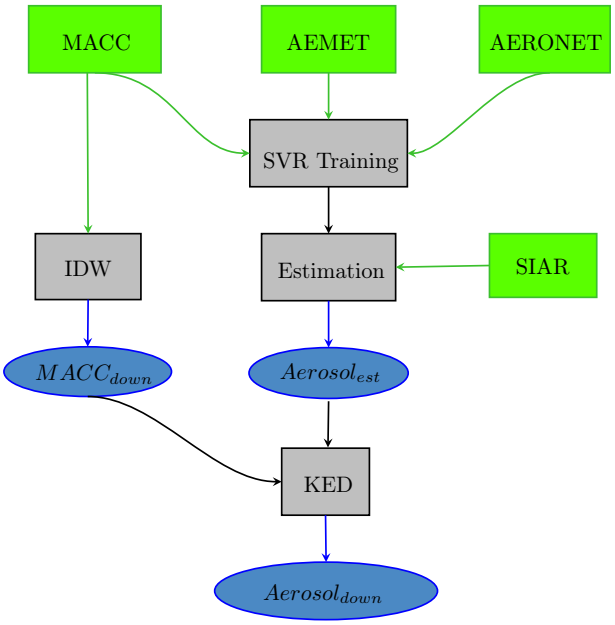

Figure 2: Methodology of the downscaling proposed. Green and grey boxes represent data sources and specific operations, respectively. Blue ellipses denote operation results.

	C	σ	ϵ
TL	0.9302	0.0904	0.1606
AOD 550	1.3575	0.1871	0.1944

Table 2: Main parameters of SVR trained from the best individual of the generation 30.

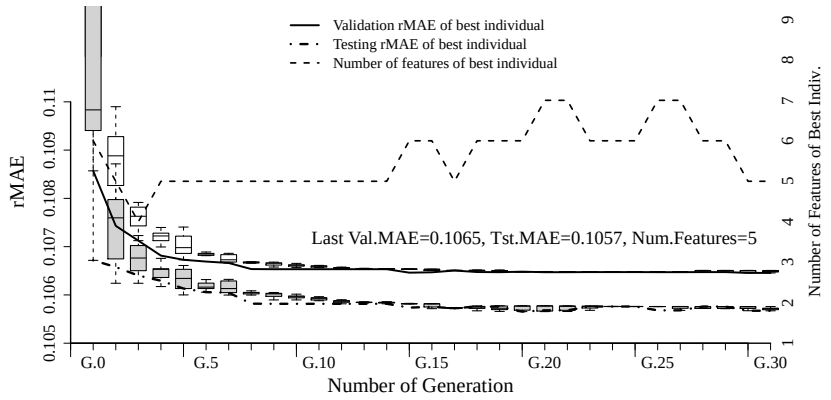

Figure 3: Evolution of validation and testing relative mean absolute errors (rMAE) of TL models along 30 generations and number of input features required in the model.

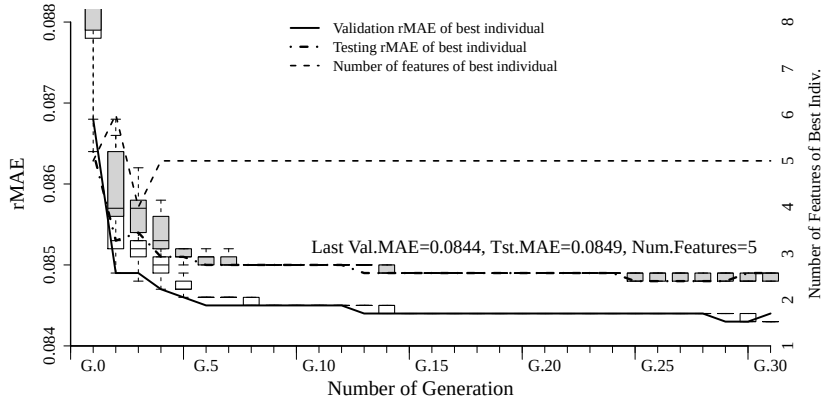

Figure 4: Evolution of validation and testing relative mean absolute errors (rMAE) of AOD 550 models along 30 generations and number of input features required in the model.

Station	$rMAE_{aer,macc}$	$R^2_{aer,macc}$	$rMAE_{down,aer}$	$R^2_{down,aer}$	$rMAE_{down,macc}$	$R^2_{down,macc}$	$rMAE_{cv,ked}$	$R^2_{cv,ked}$
1	0.174	0.479	0.091	0.819	0.167	0.547	0.123	0.683
2	0.163	0.381	0.094	0.894	0.127	0.449	0.112	0.601
3	0.186	0.175	0.070	0.768	0.120	0.219	0.184	0.514
4	0.183	0.346	0.113	0.789	0.196	0.389	0.178	0.552
5	0.207	0.444	0.116	0.819	0.117	0.465	0.083	0.421
6	0.184	0.255	0.119	0.689	0.220	0.319	0.203	0.302
7	0.198	0.321	0.107	0.822	0.198	0.378	0.122	0.489
mean	**0.185**	**0.343**	**0.102**	**0.800**	**0.164**	**0.395**	**0.143**	**0.509**

Table 3: Relative mean absolute errors and coefficients of determination (R^2) evaluated in AERONET positions for TL *aerosol variable* between AERONET and MACC databases ($rMAE_{aer,macc}$, $R^2_{aer,macc}$), downscaling proposed and AERONET ($rMAE_{down,aer}$, $R^2_{down,aer}$), downscaling proposed and MACC ($rMAE_{down,macc}$, $R^2_{down,macc}$) and cross-validation of *KED* ($rMAE_{cv,ked}$, $R^2_{cv,ked}$) in per units.

Station	$rMAE_{aer,macc}$	$R^2_{aer,macc}$	$rMAE_{down,aer}$	$R^2_{down,aer}$	$rMAE_{down,macc}$	$R^2_{down,macc}$	$rMAE_{cv,ked}$	$R^2_{cv,ked}$
1	0.584	0.194	0.087	0.645	0.357	0.244	0.148	0.319
2	0.757	0.011	0.076	0.533	0.413	0.100	0.147	0.229
3	0.611	0.023	0.054	0.493	0.399	0.084	0.137	0.368
4	0.680	0.063	0.082	0.553	0.228	0.199	0.201	0.298
5	0.640	0.024	0.119	0.620	0.327	0.173	0.132	0.530
6	0.632	0.084	0.098	0.459	0.448	0.217	0.230	0.179
7	0.690	0.083	0.087	0.555	0.403	0.193	0.099	0.277
mean	**0.656**	**0.069**	**0.086**	**0.551**	**0.368**	**0.173**	**0.156**	**0.314**

Table 4: Relative mean absolute errors (rMAE) and coefficients of determination (R^2) evaluated in AERONET positions for AOD 550 *aerosol variable* between AERONET and MACC databases ($rMAE_{aer,macc}$, $R^2_{aer,macc}$), downscaling proposed and AERONET ($rMAE_{down,aer}$, $R^2_{down,aer}$), downscaling proposed and MACC ($rMAE_{down,macc}$, $R^2_{down,macc}$) and cross-validation of *KED* ($rMAE_{cv,ked}$, $R^2_{cv,ked}$) in per units.

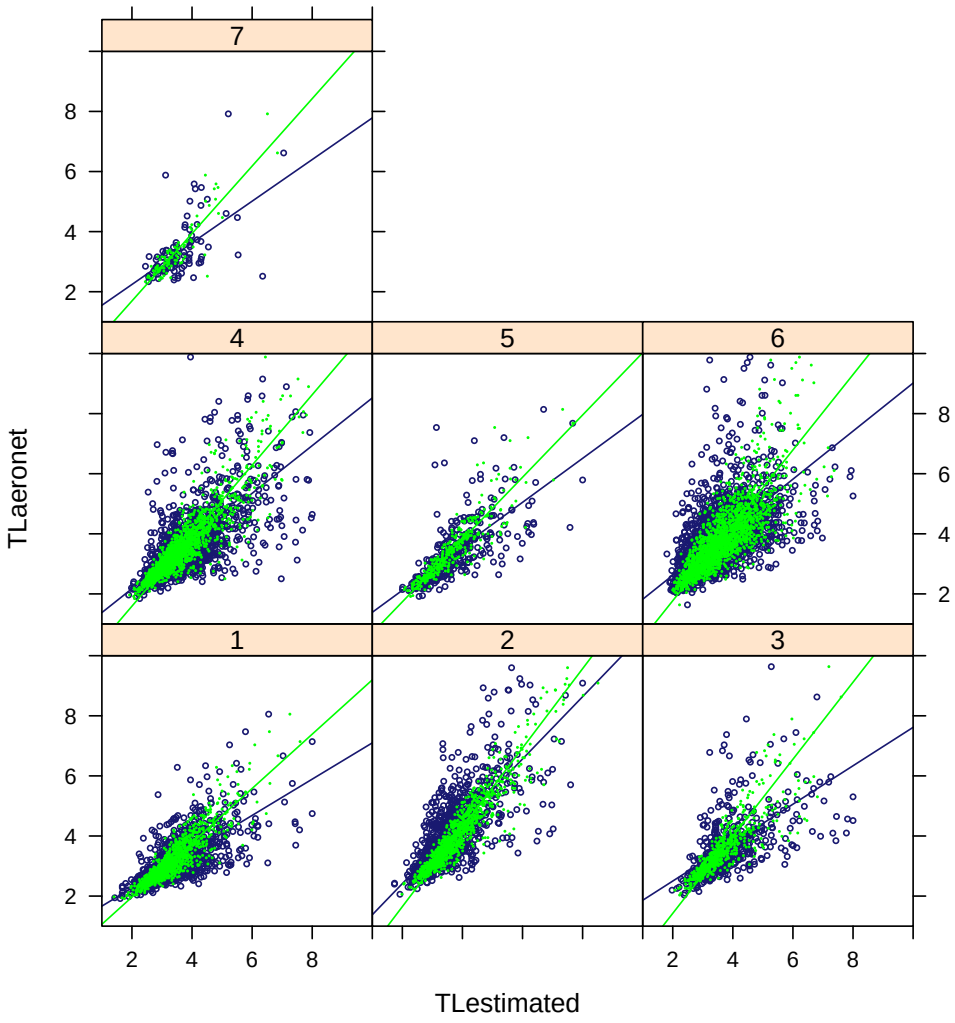

Figure 5: Scatter plots of TL from AERONET and estimated (MACC and SVR) for the seven AERONET locations. Blue points stand for MACC and green for SVR.

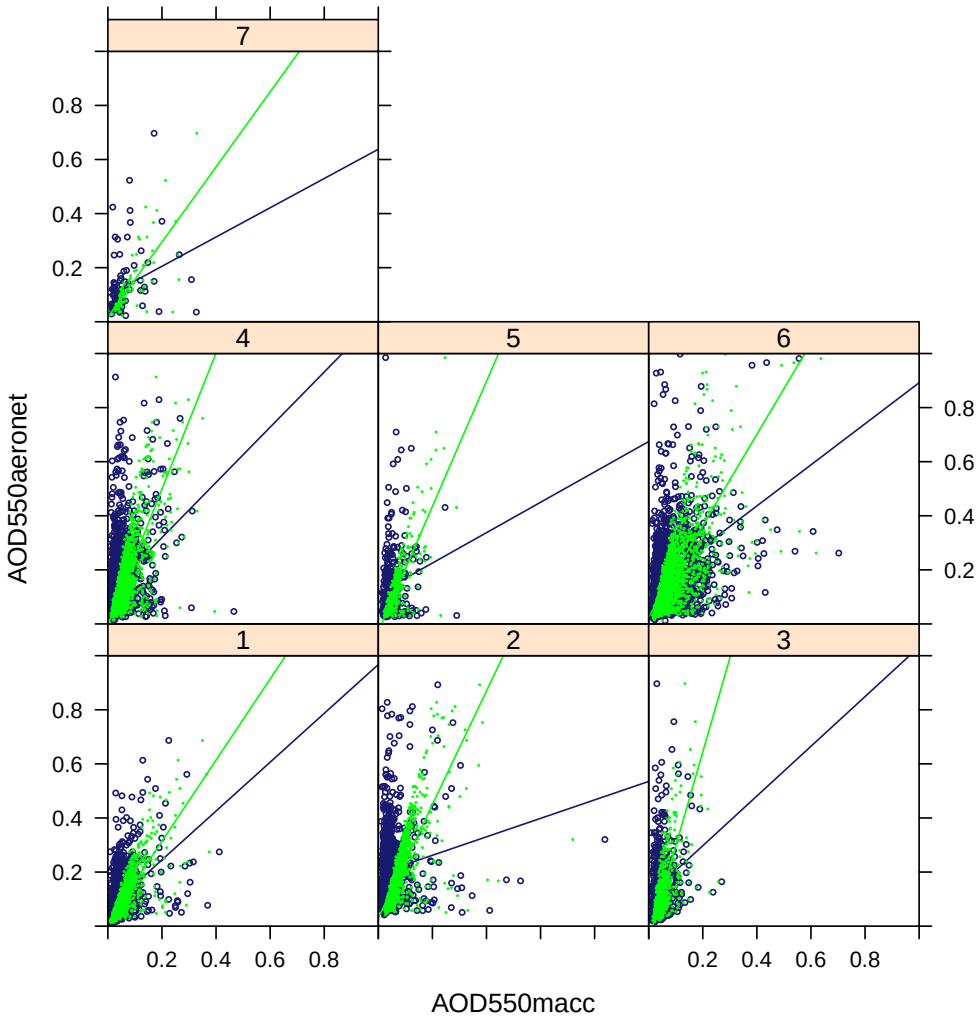

Figure 6: Scatter plots of AOD 550 from AERONET and estimated (MACC and SVR) for the seven AERONET locations. Blue points stand for MACC and green for SVR.

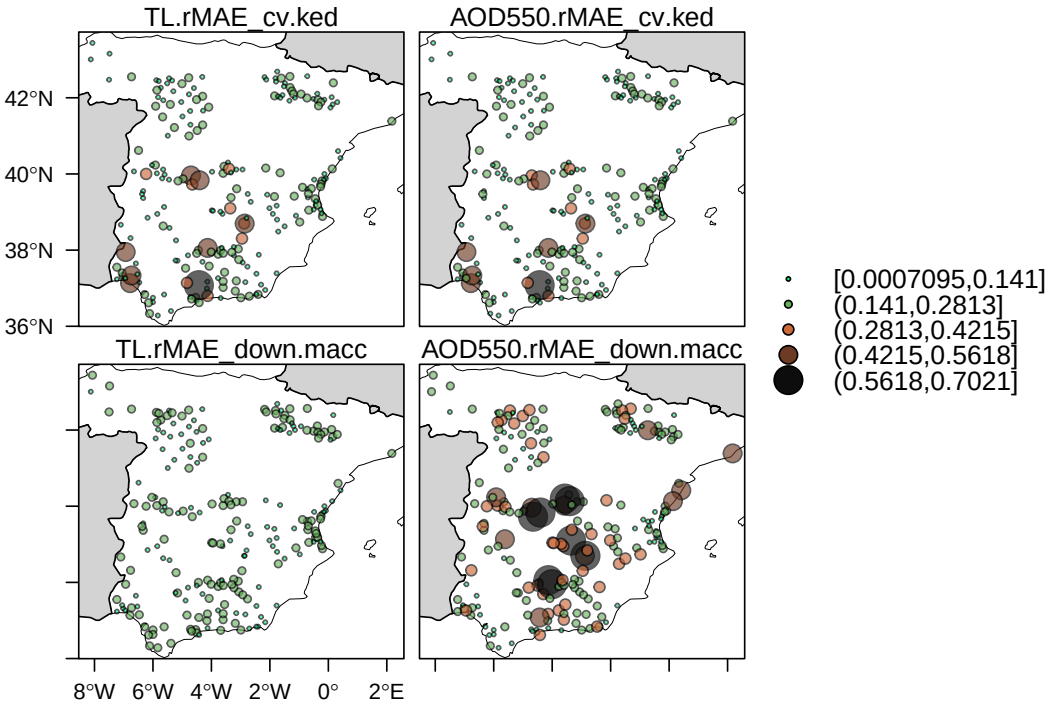

Figure 7: Bubble plots with cross-validation errors of *KED* (upper images) and $rMAE_{down,macc}$ (lower images) for TL (left) and AOD 550 (right), respectively. Each color or size corresponds to each interval of errors represented in the scale.

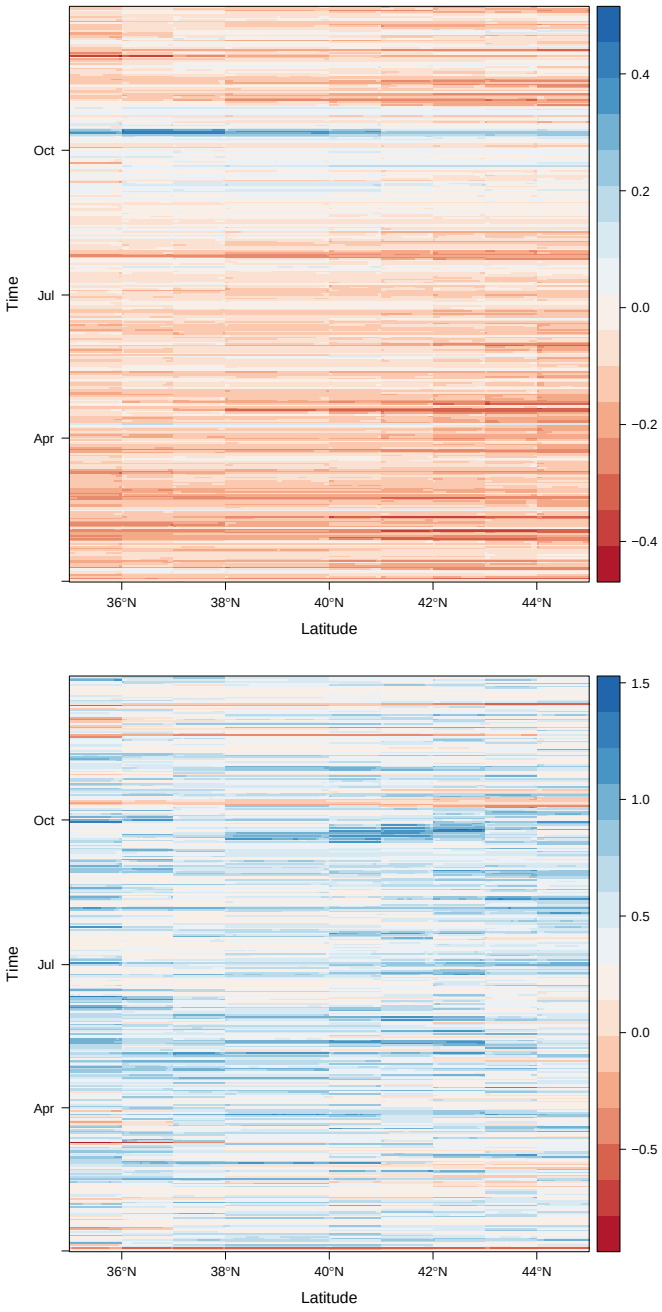

Figure 8: Hovmöller plots of the relative difference of aerosol variables related to MACC in the year 2008; TL (upper image), AOD 550 (lower image).

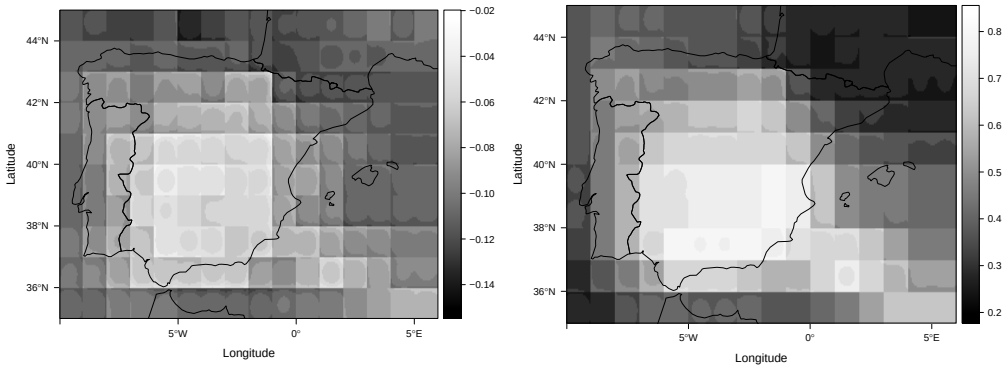

Figure 9: Annual relative difference of aerosol variables related to MACC in the year 2008; TL (left image) ($\frac{TL_{down}-TL_{macc}}{TL_{macc}}$), AOD 550 (right image) ($\frac{AOD550_{down}-AOD550_{macc}}{AOD550_{macc}}$).

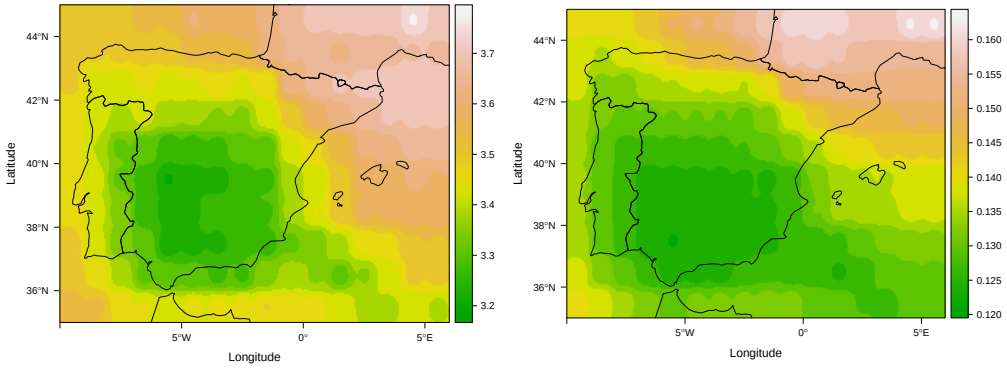

Figure 10: Annual mean aerosol variables in the year 2008; TL (left image), AOD 550 (right image).