Royal Netherlands
Meteorological Institute
Ministry of Infrastructure and the Environment

Deriving erythemal UV dose rate from Meteosat Observations

E. Wiegant and J.F. Meirink

De Bilt, 2017 | Technical report; TR-359

Deriving erythemal UV dose rate from Meteosat observations

Evert Wiegant
Jan Fokke Meirink

September 5, 2016 - January 6, 2017

Abstract

Downwelling solar irradiance at the surface is retrieved from the geostationary Meteosat SEVIRI instrument using the SICCS algorithm. The existing SICCS broad-band irradiance product is being extended with irradiance in 5 separate wavelength bands, including two in the ultraviolet (UV). In this study, the first band (283 - 328 nm) has been validated using various Brewer measurements across Europe using data from January and July of 2014. Next, it is investigated whether the improved spectral irradiance information allows estimating the erythemal UV dose rate (UVDr). The UVDr is computed from the Brewer measurements and compared to various SICCS output parameters and cloud physical properties in order to construct a fit to compute the UVDr. The best UVDr model results from fitting clear sky SICCS band 1 irradiance to the ground-based UVDr, to which a cloud correction is applied that is based on cloud optical thickness.

Chapter 1

Introduction

1.1 Erythemal UV

Ultraviolet (UV) light, although not visible to humans, has various effects on life on earth. A well known effect of UV light on humans is tanning, or, when overexposed, burning of the skin. The contribution of UV light to burning of the skin for each wavelength λ is described by the erythemal action spectrum $E(\lambda)$ (McKinlay & Diffey 1987):

$$\begin{array}{lll} E(\lambda) = 1 & \text{for} & \lambda \leq 298\text{nm} \\ E(\lambda) = 10^{0.094 \cdot (298 - \lambda/\text{nm})} & \text{for} & 298\text{nm} < \lambda \leq 328\text{nm} \\ E(\lambda) = 10^{0.015 \cdot (139 - \lambda/\text{nm})} & \text{for} & 328\text{nm} < \lambda \leq 400\text{nm} \\ E(\lambda) = 0 & \text{for} & 400\text{nm} < \lambda \end{array}$$

This action spectrum, also visualized in figure 1.1, is used to compute the UV index (UVI), an international standard used primarily to communicate information on the amount of UV on a certain day to the general public (Vanicek et al. 2000, World Health Organization 2002). The UVI of a particular day is defined as the amount of effective irradiance for erythema during solar noon under clear sky and is computed as follows:

$$UVI = \frac{1}{25\text{mW/m}^2} \int S_{noon}(\lambda)\, E(\lambda)\, d\lambda \tag{1.1}$$

Here $S_{noon}(\lambda)$ is the solar irradiance under clear sky at the earth's surface at solar noon. Note that the UVI is a unitless quantity. Similarly we can use the erythemal action spectrum to compute the actual effective irradiance (at any moment for any sky type) which we call the UV dose rate (UVDr):

$$I_{UVDr} = \int S(\lambda)\, E(\lambda)\, d\lambda \tag{1.2}$$

Where I_{UVDr} is the erythemal effective irradiance (the UV dose rate).

SICCS band	Wavelengths
1	283- 328 nm
2	328- 408 nm
3	408- 704 nm
4	704-1194 nm
5	1194-4606 nm

Table 1.1: Wavelength bands for which the SICCS algorithm calculates surface irradiance.

Integrating all UV dose rates throughout the day yields the UV dose (UVD), which is another quantity found in literature as a UV proxy (often expressed in kJ/m^2):

$$I_{UVD} = \int_{day} \left(\int S(\lambda, t)\, E(\lambda)\; d\lambda \right) dt \tag{1.3}$$

1.2 SICCS algorithm

The SEVIRI instrument aboard the geostationary MSG satellites makes high resolution (3x3 km at nadir) scans of the full disc of the earth (about 0,0-latitude/longitude) every 15 minutes. These scans are used to derive cloud physical properties. The SICCS (Surface Insolation under Clear and Cloudy skies derived from SEVIRI imagery) algorithm (Greuell et al. 2013) computes downward surface radiation using these cloud physical properties and other quantities from different data sets (e.g. aerosol and ozone data). The present version computes broad-band surface insolation, which will be extended in the future with insolation in separate wavelength bands. These bands include the wavelengths shown in table 1.1.

1.3 Objectives

The first objective of this study is to validate the SICCS band 1 output using ground-based measurements using Brewer instruments. The broad-band version of the SICCS product has been validated with ground-based BSRN stations (Greuell et al. 2013). The new version requires validation using ground-based measurements of the separate bands. Brewer measurements are a good candidate for validation of band 1. Many institutes worldwide make Brewer measurements, so we should be able to collect a good data set. Validation of band 1 is done in chapter 3.

The second objective of this study is to derive the UVDr from SICCS input and/or output. The first two bands of the SICCS algorithm capture all the wavelengths that are important for the erythemal action spectrum. This is illustrated in figure 1.1, together with the wavelength ranges of some of the spectrometers used in this study (Brewer and Bentham spectrometer). We therefore

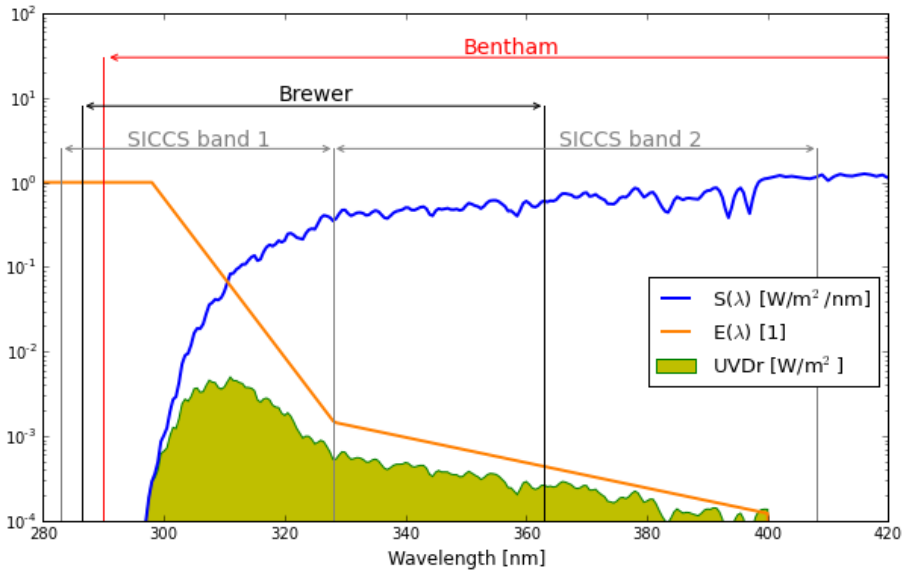


Figure 1.1: A typical solar spectrum at the earth's surface (blue), the erythemal action spectrum (orange) and the resulting UVDr (yellow). The wavelength ranges of SICCS bands 1 and 2 and the measuring ranges of the Brewer and Bentham spectrometers are also indicated.

think that we have sufficient information (using the cloud physical properties and the SICCS irradiances) to make a good estimate of the UVDr. Ground-based data of the UVDr can be derived from the same Brewer measurements that we use to validate band 1. We will study how to estimate the UVDr in chapter 4.

Chapter 2

Data & methods

The SICCS version including the separate bands only has a small archive. Two months (January and July) of 2014 are available for our study.

Both objectives require comparison with ground-based measurements of spectra in the UV range. The ground-based data that are used for this study are shown in table 2.1.

We mostly have data from Brewer spectrometers. We also use data from one Bentham spectrometer and two spectrometers from which the erythemal irradiance is directly deduced. These are a NILU instrument and YES pyrannometer in Thessaloniki. The Bentham spectrometer has a wavelength range that is ideal for the UVDr and both SICCS bands 1 and 2 as it measures wavelengths from 290 - 500nm. The Brewer measurements only include measurements from 286.5 up to 363nm. These wavelength ranges in comparison to a typical irradiance spectrum and the UVDr are shown in figure 1.1. The Brewer measurements do not span the entire range of wavelengths that are important for the erythemal UVDr. We need to correct for this discrepancy somehow in order to calculate the UVDr from Brewer measurements. This will be discussed in section 2.1. We will discuss the data selections we use for the data we decide to study in

Location	Instrument	Lat [°]	Lon [°]
Reading, GB	Bentham	51.44	-0.94
De Bilt, NL (KNMI)	Brewer	52.10	5.18
Bilthoven, NL (RIVM)	Brewer	52.12	5.20
Uccle, BE	Brewer	50.80	4.36
Thessaloniki, GR	Brewer	40.96	22.96
Thessaloniki, GR	NILU & YES	40.96	22.96
Paramaribo, SR	Brewer	5.81	-55.21

Table 2.1: Ground-based measurements available for our study. Note that the Bentham instrument in Reading is operated by the university of Manchester, and will be referred to as either 'Reading' or 'Manchester' in this report.

section 2.2. Data selection is needed because the ground measurements do not perfectly correspond to SEVIRI measurements in time and place.

2.1 UVDr correction

The UVI or UVDr are commonly derived from Brewer measurements. The problem of correcting for the limited Brewer spectrum has been solved before by Marc Allaart (reference: private communications). The correction is done by giving the action spectrum a lower limit of $1.11 \cdot 10^{-3}$, then integrating over the available wavelength range (up to 363 nm). This correction is expected to work well because both the part of the solar spectrum to which this correction is applied, as well as the part that is not measured by the Brewer spectrometer (363-400nm), are least sensitive to ozone.

This method is tested using the Bentham data, which has wavelengths available up to 500nm. We computed the UVDr from the Bentham data, and compared it to a UVDr that is integrated to 363nm with the uncorrected action spectrum (figure 2.1a) and the UVDr integrated to 363nm with the corrected action spectrum (figure 2.1b). We can see that for both comparisons, the correlation is very close to one, but the uncorrected UVDr is biased. This bias may be corrected with a constant factor, but this is discouraged because the corrected action spectrum yields a correlation that lies about 5 times closer to 1 than the correlation of the uncorrected action spectrum:

$$\frac{1 - r^2_{uncorrected}}{1 - r^2_{corrected}} \approx 5$$

For the rest of our studies we use the corrected action spectrum for calculating any UVDr from ground-based spectral measurements.

2.2 Data selection

The SEVIRI instruments scans the earth every 15 minutes at regular intervals. The ground-based measurements may not be as regular and/or be performed at different times. We describe the general data selection method here, and discuss how exceptions from this method are handled.

2.2.1 Matching ground data with SICCS data

The data selection for most chapters is similar. For each ground-based insolation value, a corresponding SICCS value is searched. The SEVIRI instrument produces one scan of the disc per quarter hour. The scan time of a certain pixel is the time of the start of the scan (every whole quarter) plus the offset for the specific location. For each ground-based measurement, the closest scan time for its corresponding SEVIRI pixel is found that lies within 7.5 minutes of the ground measurement time. If the SICCS data set is complete, this means

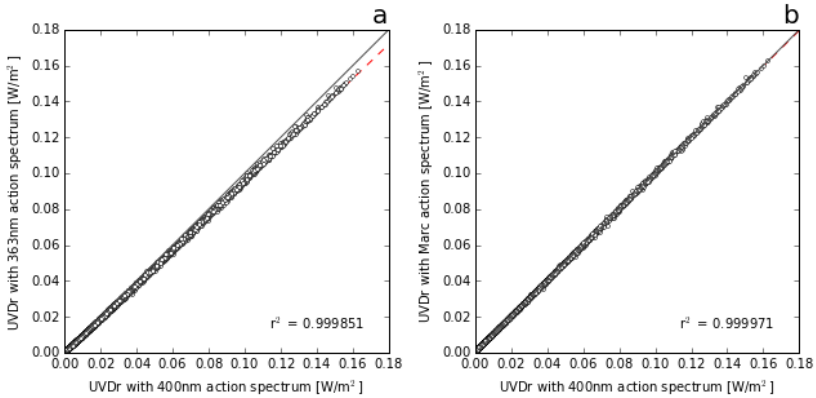


Figure 2.1: Comparison of the UVDr computed from the Bentham spectral measurements of July 2014, using the full (400nm) erythemal action spectrum and the uncorrected (a) and corrected (b) 363nm limited action spectrum. The grey line shows the 1:1 line and the least squares fit is shown as the dashed red line.

that every ground measurement has a corresponding SICCS insolation value. If a quarter is missing and the closest scan time is more than 7.5 minutes away from the ground measurement time, the ground measurement is discarded. No distinction is made between measurements performed later than or earlier than the corresponding SICCS value.

2.2.2 Exceptions

Measurement frequency

Different instruments measure at different times and frequencies. The NILU and YES instruments measure every minute, which is much more frequent than the Brewer/Bentham spectrometers. SEVIRI measurements are therefore not selected when they lie within 7.5 minutes of these measurements, but only if they lie within 30 seconds. This prevents the same SEVIRI measurement from being selected for multiple ground measurements.

Comparison between KNMI & RIVM

Ground measurements of both KNMI and RIVM are selected as described in section 2.2.1, but must also correspond to each other when compared in section 3.3. The RIVM measured every 12 minutes, so for every KNMI measurement, a RIVM measurement is sought that is performed within 6 minutes of the KNMI measurement. If there is no RIVM measurement within 6 minutes of a specific KNMI measurement, the KNMI measurement is discarded. All RIVM measurement that were not selected by this procedure are also discarded. If a RIVM

measurement is selected by 2 KNMI measurements, the last KNMI measurement is discarded.

Weight

The selection method is designed to use as many ground-based measurements as possible. Because some stations measure more frequently than other stations, the amount of data points of different stations will not be equal. If each point is given the same weight, the model will be biased towards predicting values for the better-represented station. To avoid this, every point should be given a weight, corresponding to the amount of points with which it shares its station.

The least squares fitting method can easily be given weights, e.g. equal to 1/(#points) for each station's data. For other fitting methods, such as the RANSAW method that is used most often, such a weight is more difficult to define. A weight could be introduced when counting inliers, such that points of better-represented stations count for less. However, this is not done, mainly because the fitting procedure is already manually adjusted by the person performing the fit, such that the result will be a good fit.

Note on quality

The Brewer instruments take a relatively long time to measure the spectrum (about 4 minutes). Cloud cover may change during such measurements. The satellite product gives single values for the cloud properties (e.g. the cloud phase). When comparing satellite and ground values we may be looking at different weather types if cloud cover changes rapidly.

Some stations provided data of high temporal resolution broadband insolation measurements from which the variation during the Brewer measurement can be deduced. The standard deviation of 5 values roughly corresponding to the time of the Brewer measurement are computed and divided by their mean. The resulting value is used as a flag.

Points with large flag values are expected to 'behave' as a different weather type than what they are classified as. Clear sky points generally correlate well with their ground measurements. Clear sky points with high flag values may be expected to lie further from the fit, or be *more likely* to lie further. However, this behavior is not found. That is, points with large flag values do not consistently lie further from the fit than points with small flag values.

We may question whether we have stated and tested the entirety of expected behaviors. However, there may be similarities in the high spatial and small temporal extent of the satellite observation, and the high temporal and low spatial extent of the Brewer measurement. A pixel with scattered cloud cover will have an average optical thickness observed by the satellite based on patches of clear and cloudy sky. The same clouds drifting over in time give the Brewer measurements an average value of the same clear and cloudy patches, albeit with a different distribution of wavelengths.

Chapter 3

Validation of SICCS UV bands

In this chapter we validate band 1 of the new SICCS version by comparison with ground measurements. Spectral UV data that includes all wavelengths of band 2 is meagerly available (only the Bentham instrument). We therefore limit ourselves to validating band 1. We will consider both months separately, starting with July in section **??** followed by January in section **??**.

The data of the KNMI and the RIVM allows us to do a study on small scale variations because of their close proximity. We compare ground-based band 1 irradiances and SICCS band 1 irradiances to see if the SICCS product shows a similar variation to the variation of the (ground-based) observations. This is done is section 3.3.

3.1 July

Scatter plots of SICCS vs. ground-based band 1 insolation for each station and sky type during July 2014 are shown in figure 3.1. The relative overall RMSE is 28% with a correlation coefficient of 0.94. SICCS especially tends to underestimate insolation in band 1 at high values. We can see that the prediction for clear sky is generally better than for cloudy skies. The relative RMSE of clear sky cases is mostly lower than 20% and somewhat over 30% for either water or ice cloud cases. Insolation at Thessaloniki is particularly well predicted for clear skies, while for cloudy skies the prediction skill is much worse than for the other stations. However, most clear sky predictions appear to underestimate the actual (ground observed) value. To get an idea of the bias involved, we will plot the difference (Δ) between the insolation of SICCS and the ground observed value as a function of an input parameter of the SICCS model/algorithm. One parameter we expect to introduce a bias for the clear sky case is the aerosol optical thickness. Aerosols are not used in computation of either cloudy case. Plotting Δ as a function of aerosol optical thickness is

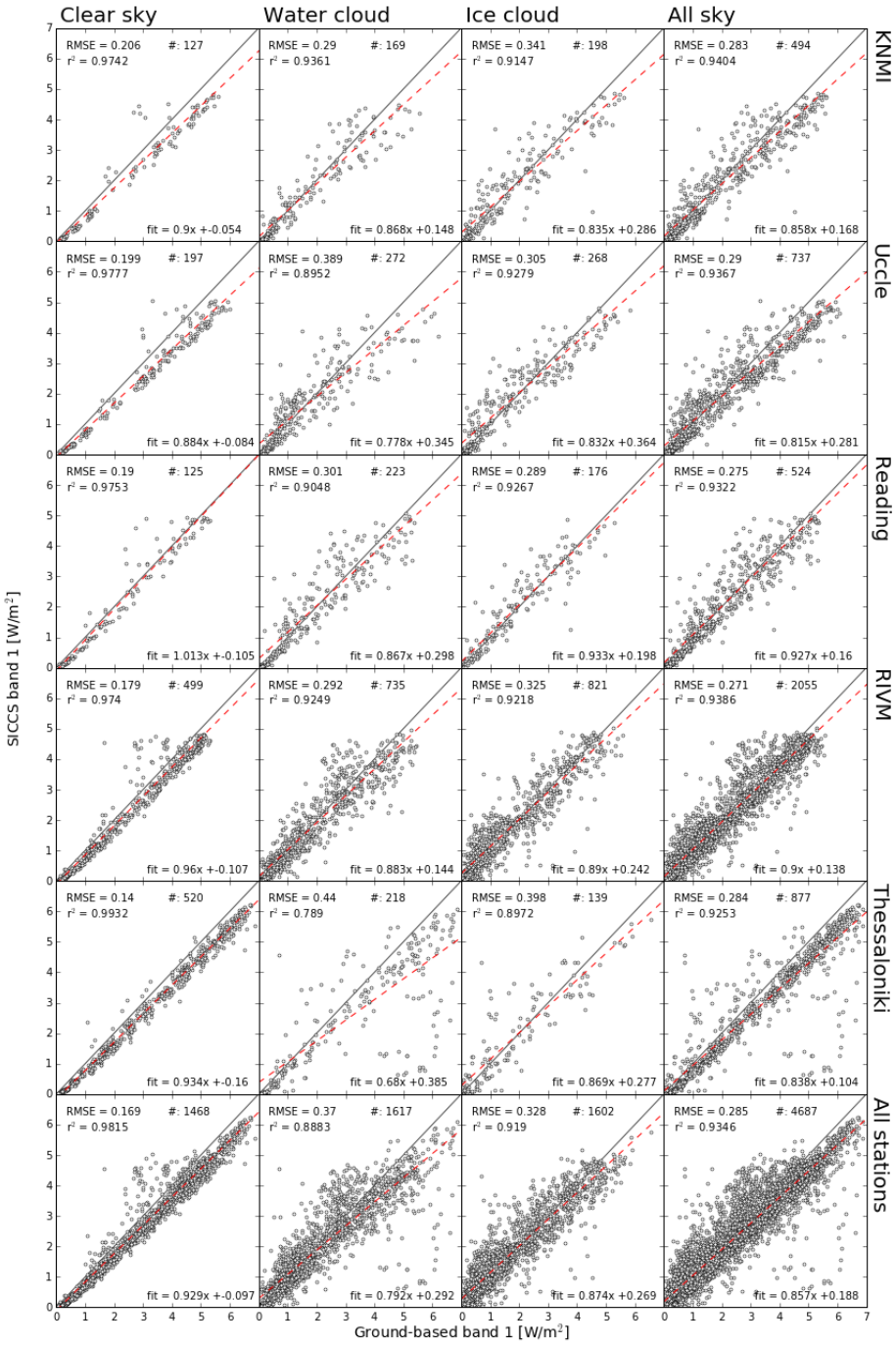


Figure 3.1: Comparison between SICCS and ground-based band 1 insolation measurements. Data of July 2014 at different stations and cloud types are shown. In each panel the 1:1 line is shown in grey and OLS fit in dashed red and in formula (fit) in the bottom right. Also shown are the root-mean-square error relative to the mean (RMSE), correlation coefficient (r^2) and number of points (#).

done in figure 3.2. Another parameter that seems to affect the bias is the water vapour path, which is shown in figure 3.3.

The aerosol optical thickness may have been expected to explain the bias. However, the bias is mostly constant for every aerosol optical thickness. We cannot be conclusive about what causes the bias as we have not found any other parameters that affects the bias (i.e. zero bias for zero of the parameter value and nonzero bias for nonzero parameter value).

The water vapour path, shown in figure 3.3, shows a surprising relation between the differences in insolation Δ. Water vapour is not expected to affect UV insolation directly, but may have indirect effects that originate from the retrieval of cloud properties. The SICCS insolation seems to increase with respect to the ground-based measurement for larger water vapour paths. This appears especially true for ice clouds, where the RIVM data is particularly convincing.

3.2 January

Scatter plots of band 1 insolation for each sky type and station are shown in figure 3.4.

Notable for band 1 is the bias. The ODR fit parameters that are shown in the bottom right corners of each panel show an underestimation of about 60% at its lowest. Clear sky points show a different relation than cloudy sky points; the linear ODR fit does not cross the origin and has a slope closer to 1. However, a small 'bend' occurs at the lowest irradiance values such that points still tend towards the origin. This is especially visible for the Reading and RIVM clear sky panels (of figure 3.4). This behavior may be present in July, but we may have missed it as we have not zoomed in so far on the lowest irradiance values. Also, the July data was saved in a coarser format when it was analyzed, from which it is difficult to see this behavior at all.

Biases in January data are also studied as was done for July data. Biases (SICCS data minus ground observed data) as a function of some parameters (aerosol optical thickness and water vapour path) are shown in figures 3.5 and 3.6. The results are similar to those of July: There is little effect from the aerosol optical thickness and the bias becomes more positive for larger water vapour paths.

3.3 Comparing RIVM and KNMI

The dutch national institute for public health and the environment (RIVM) and the Royal Netherlands Meteorological Institute (KNMI) are located in close proximity to each other. Both institutes have performed measurements of UV spectra. Having two independent data sets of gives us the opportunity to study variation in insolation on a small spatial scale. We can compare this to the SICCS-computed insolation.

Comparing the ground stations' data in a scatter plot (figure 3.7a) we find

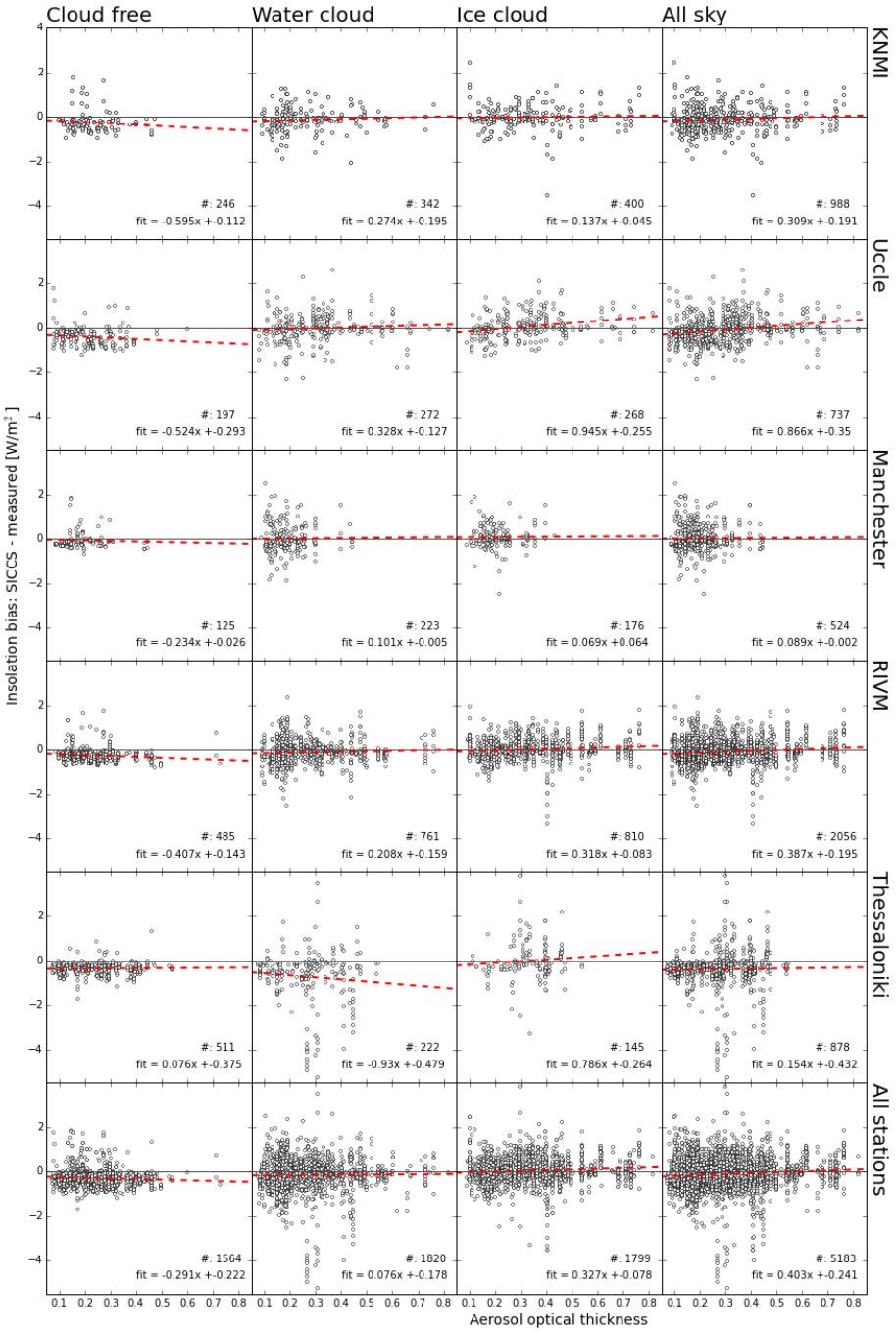


Figure 3.2: The difference (Δ) between SICCS band 1 insolation and corresponding ground observations of July 2014, as a function of the aerosol optical thickness. The grey line shows zero bias, the dotted red line is the least squares fit which is also shown in formula (fit). The number of points (#) is also displayed in each panel.

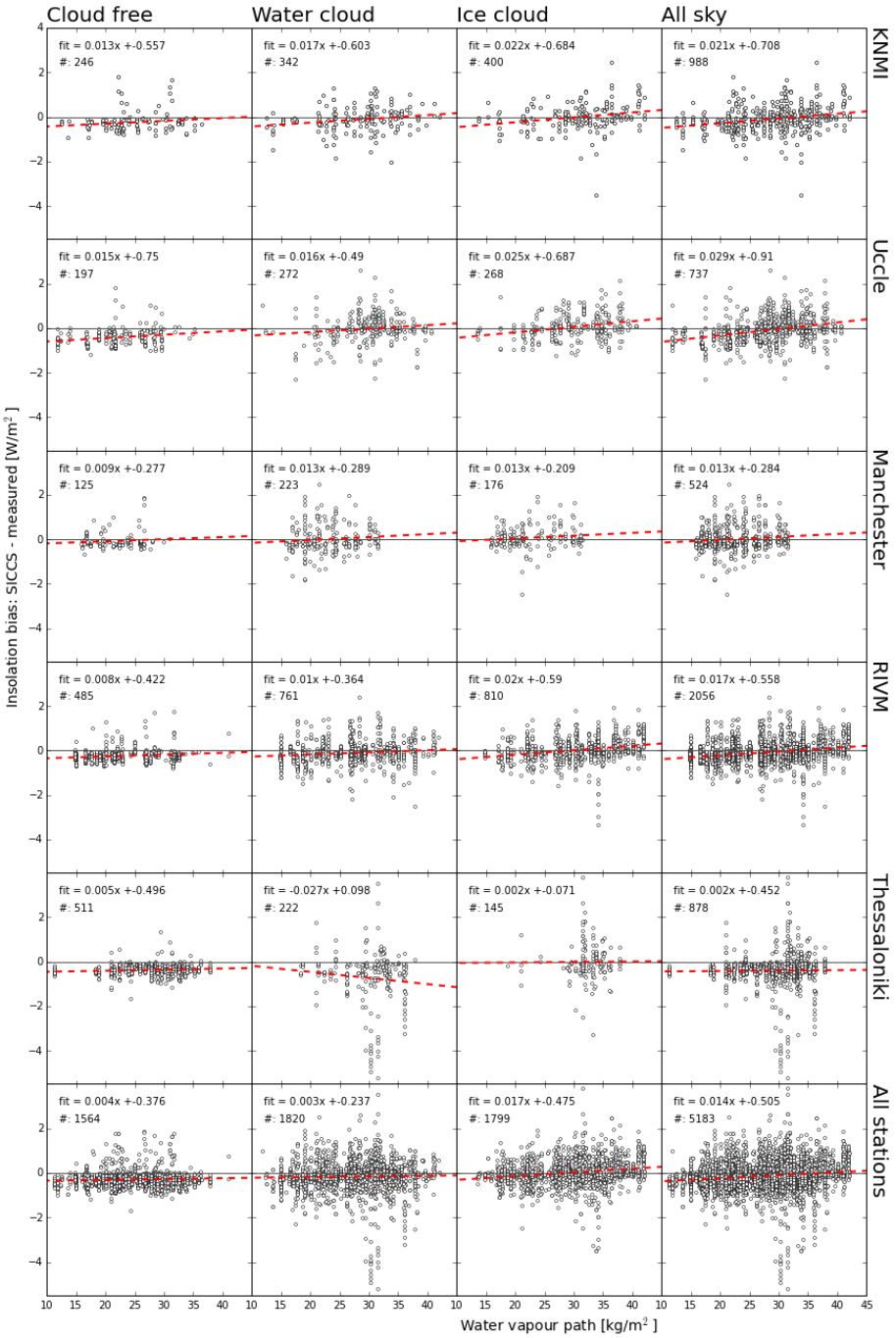


Figure 3.3: The difference (Δ) between SICCS insolation from band 1 and corresponding ground observations, as a function of the water vapour path. The grey line shows zero bias, the dotted red line is the least squares fit which is also shown in formula (fit). The number of points (#) is also displayed in each panel.

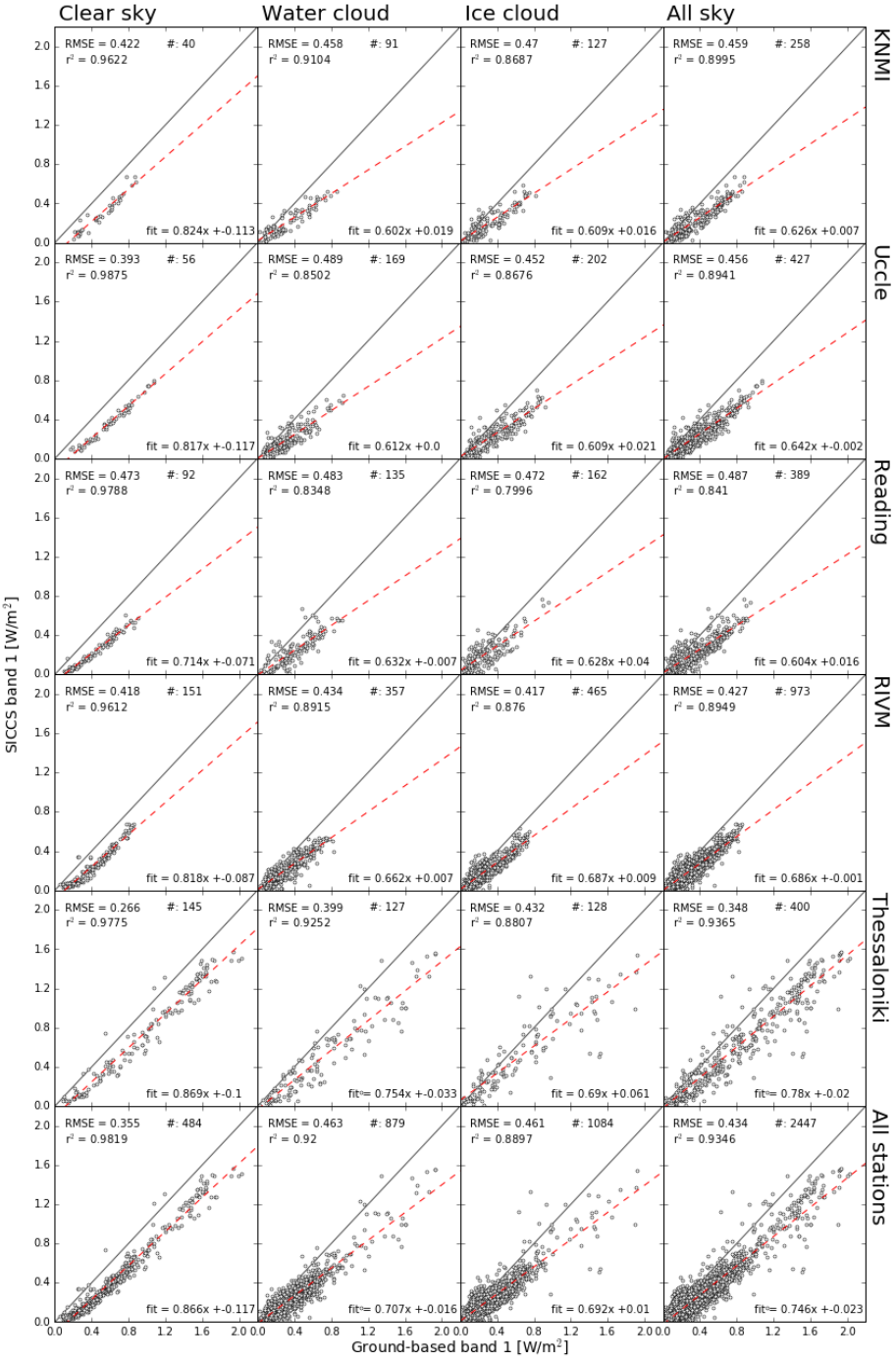


Figure 3.4: As for figure 3.1 but for January 2014.

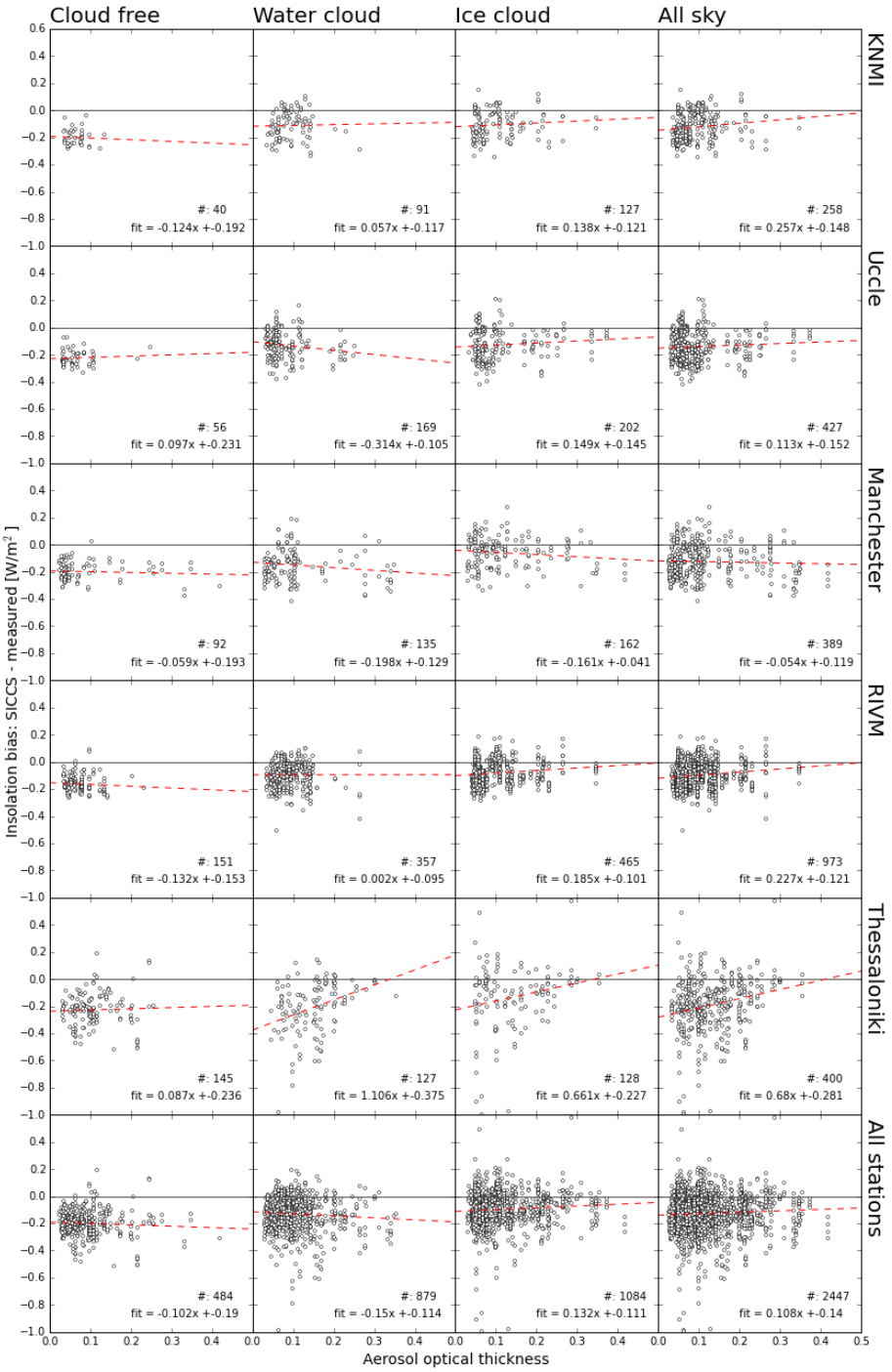


Figure 3.5: As for figure 3.2 but for January 2014.

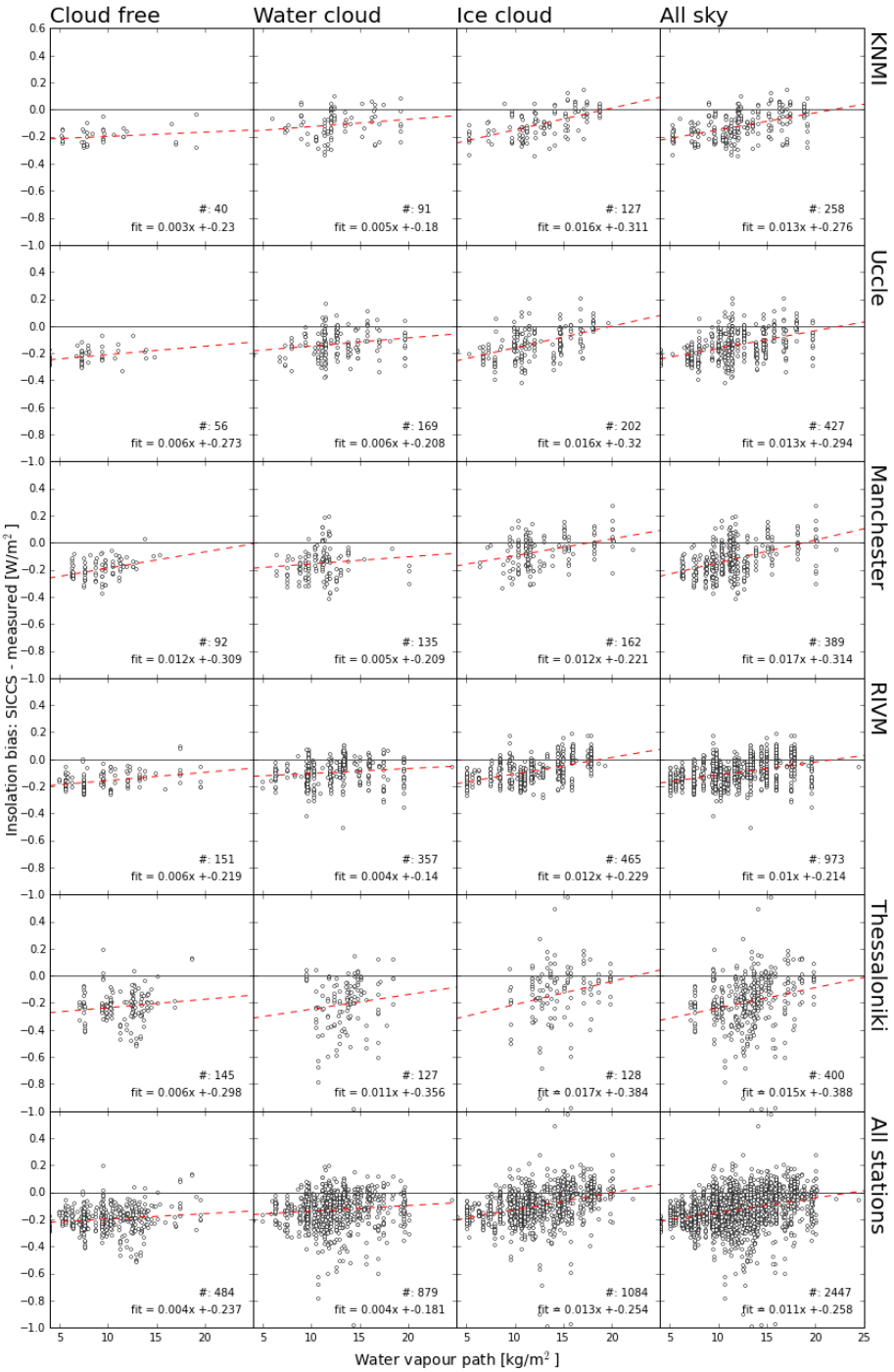


Figure 3.6: As for figure 3.3 but for January 2014.

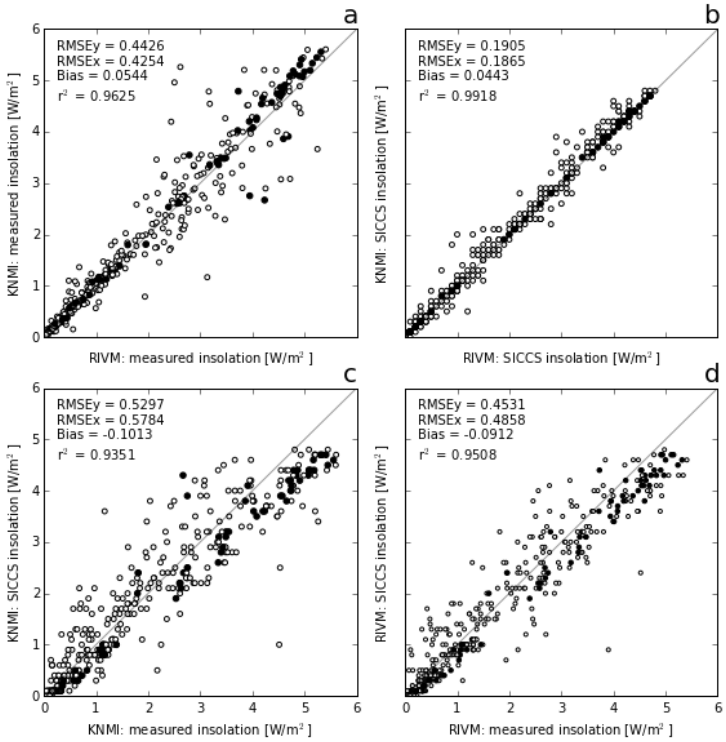


Figure 3.7: Comparison of KNMI and RIVM ground measurements (a) and the SICCS insolation values of their corresponding SEVIRI pixels (b). Ground measurements of each station is also compared to the SICCS insolation values of its representative pixel (c for KNMI, d for RIVM). Black dots are for clear sky, white dots for cloudy.

there is some degree of agreement, but also significant spread. As there is little bias, we assume the ground stations to be reliable and that the spread is solely attributable to spatial variation (i.e. if the stations where right next to each other they would measure the same values).

The ground measurements are also compared to their corresponding SICCS values in figures 3.7b and c, and the SICCS values corresponding to each station are compared in figure 3.7d . The comparison between both stations' SICCS values shows the degree of variability that the model predicts. We can see that this degree of variability (i.e. the amount of spread) is much smaller than the variability of the ground stations' measurements. We can therefore question whether the SICCS model is able to resolve variability on the small scale on which SEVIRI does its observations. We can attribute this smaller spread in the model values to the representation problem; the expected discrepancy between a point value and a value representative for a grid box spanning multiple km^2.

However, it is well possible that the model is simply not accurate enough to represent the ground values, and thus not accurate enough to resolve fine-scale variations. We can deduce this from the fact that the ground measurement do not agree too well their corresponding SICCS values. The data from one station in fact correlate better to the data of the other's SICCS value (correlation table 3.1). However, I would rather say that it does not matter which pixel represents which station, as their correlations are all similar. Then, because it does not matter which pixel represents which station, it also does not matter whether the model computes on the current scale, or a coarser scale.

		Ground-based		Satellite observed	
		KNMI	RIVM	KNMI	RIVM
Ground-based	KNMI	1	.9625	.9351	.9411
	RIVM	·	1	.9442	.9508
Satellite observed	KNMI	·	·	1	.9918
	RIVM	·	·	·	1

Table 3.1: Correlations of all quantities.

Summary: Adjacent pixels agree too well for their values either to be credible regarding accuracy, or regarding ability to resolve fine scale variation. The discrepancy between satellite observation and ground-based measurements cannot be attributed to the representation problem, simply because the representation of variation that the model DOES provide is not in agreement with observations. Computing on a smaller scale is expected to yield less spread yet, instead of more between adjacent pixels.

Chapter 4

UV dose rate from SICCS

In this chapter we aim to derive the UVDr from SICCS input (cloud physical properties) and output (irradiance in separate bands).

4.1 Relate SICCS input/output to UVDr using second-order polynomial fit

Scatter plots of the UVDr vs. SICCS band 1 & 2 irradiances are shown in figure 4.1. There is a non-linear relation between the UVDr and either band 1 or 2. We first attempted a second order polynomial fit that includes all data of both bands. Multivariate ordinary least squares (OLS) regression is used with the following fit:

$$I_{UVDr} = C_1 + C_2 \cdot I_1 + C_3 \cdot I_2 + C_4 \cdot I_1 I_2 + C_5 \cdot I_1^2 + C_6 \cdot I_2^2 \tag{4.1}$$

where I_1 and I_2 represent SICCS insolation from bands 1 and 2 respectively. The resulting fit parameters are found in table 4.1.

$C_1 =$	0.00397139
$C_2 =$	0.03392684
$C_3 =$	0.00147897
$C_4 =$	0.0033724
$C_5 =$	0.0217723
$C_6 =$	0.00014971

Table 4.1: Fit parameters for the fit in equation 4.1 using all available ground-based measurements from KNMI, RIVM, Reading, Uccle and Thessaloniki of July 2014.

To get an idea of how well the model predicts values outside the temporal/spatial range it was fitted to, we choose to train the model on data of one station and use it to compute values at another station. We construct the model

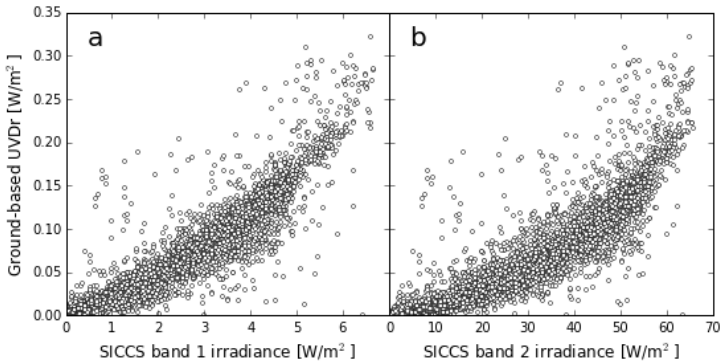


Figure 4.1: Scatter plots of the UVDr versus SICCS band 1 (panel a) and band 2 (panel b) irradiance of July 2014.

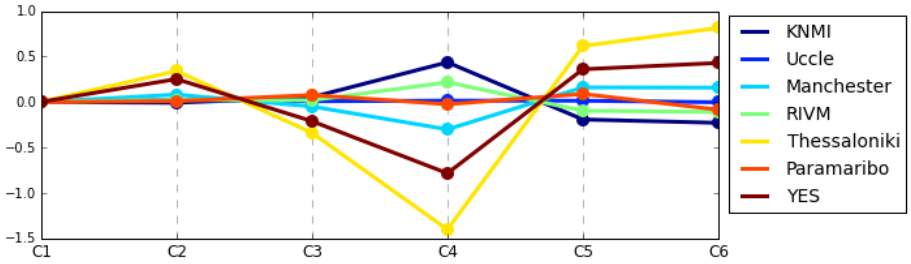


Figure 4.2: The 6 model parameters resulting from training the model using data of single stations. Each parameter is multiplied by the mean of all data of the respective insolation or product of insolations such that parameters of large quantities are not dwarfed.

for each data set, with each model consisting of the 6 parameters of equation 4.1. A graph of these parameters for each station is shown in figure 4.2.

The parameters appear to be widely varying, with some stations returning completely opposite parameter values from other stations. We therefore do not trust a model trained on one station to predict values from another station, or another location in general. In the next couple of section we explore and discuss possible solutions.

4.2 Obtaining a more robust UVDr fit model

Here is a list of possible solutions for the problem of that the UVDr fit of section 4.1 is not very robust. The final solution may be a combination of solutions.

- Make fits for each division of subsets selected by meteorological parameters (e.g. divisions by SZA, cloud type).

- Changing the model equation to include only band 1, or include SZA, COT, or other variables.
- Use a different fitting method. Orthogonal distance regression, Theil-Sen, RANSAC, etc.
- Fit to ground-based irradiance data instead of SICCS irradiance, then translate (fit) ground-based irradiance to SICCS irradiance.

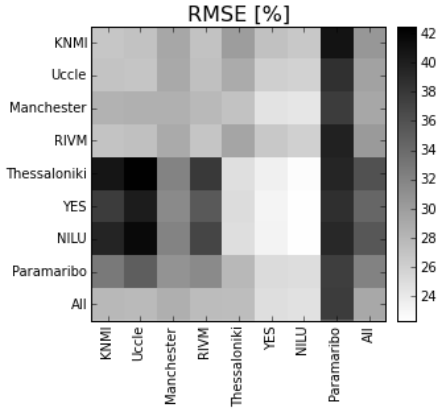
In this section we will study and/or describe how each possible solution improves the fitting method. Each of these solutions is qualified by how well it can predict values at one station using the model from another station. The root mean square error (RMSE) is generally a good measure for quality when using ordinary least squares fitting. Be careful when interpreting the RMSE for other fitting methods (e.g. Theil-Sen is deliberately used when it is NOT desirable to minimize residuals).

4.2.1 Different subsets

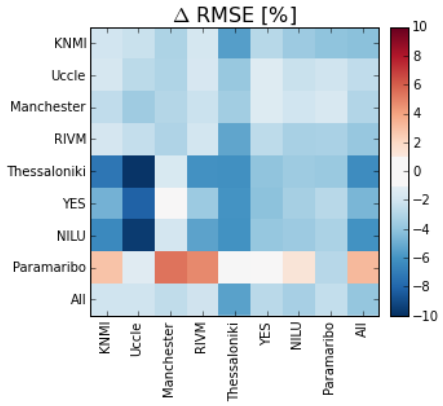
We can divide the total data set into subsets selected by parameters that we expect to affect the UVDr and/or its relation to band 1 and/or 2. We then check whether this division into subsets has improved the predictability at other stations. This check is done by comparing the RMSE for each training/testing station combination as follows: A fit of equation 4.1 is computed using SICCS band 1 and 2 (x_a), and the ground-based UVDr (y_a), using data from some station 'A'. The SICCS band 1 and 2 (x_b) data of another station 'B' is then plugged into this fit, yielding a UVDr (y'_b). This UVDr (y'_b) is then compared to the actual UVDr of that station (y_b): A linear fit is constructed with y_b (y-axis) as a function of y'_b (x-axis). The residuals in the y-direction are squared, averaged, square-rooted, then *divided by the mean of* y_b and multiplied by 100 such that the result is a percentage. The fits are made using ODR (instead of OLS) regression because python's OLS module was not able to find fits for certain subsets, whereas python's ODR module was able to do this.

The data sets (i.e. stations) on which the model is trained and tested include all available UV data sets, thus including the non-Brewer measurements from Thessaloniki. Another data set is added for training and testing which includes multiple stations at once, called the 'All' station. 'All' only includes Brewer data of KNMI, Uccle, Reading, RIVM, and Thessaloniki (so Paramaribo and Thessaloniki's YES and NILU are *excluded*).

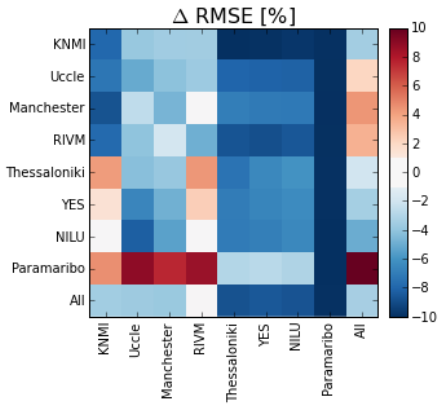
For data divided in multiple subsets (i.e. divided using criteria based on some meteorological parameter), the RMSE is computed including only data from each subset. A 'total' RMSE is then computed from these RMSEs by averaging them, weighted by the amount of points in each subsets. This total RMSE is a good measure for the overall effect of the specific division of subsets, and is compared to a reference RMSE. This reference RMSE is computed for no division into subsets and is shown in figure 4.3a.



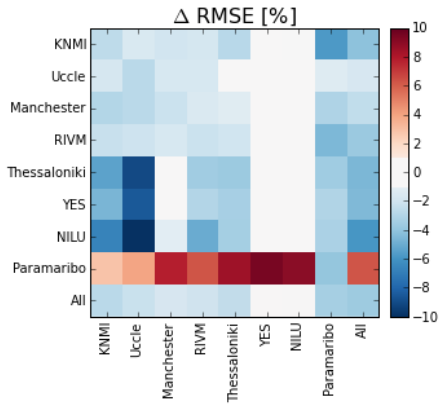
(a) The reference RMSE.



(b) Division by cloud type.



(c) Division by solar zenith angle.



(d) Division by the fraction of direct light.

Figure 4.3: The reference RMSE for each training/testing station combination and weighted averaged RMSE when dividing the data into subsets minus the reference RMSE. All fitting is performed using model equation 4.1. Training stations are on the y-axis and testing stations on the x-axis.

The number of subsets for a certain division (e.g. taking 9 ranges of SZA, or 3 cloud types, etc.) affects the resulting total RMSE regardless of the meteorological quantity by which it is selected. A randomly chosen division of 10 subsets yields a smaller RMSE than if 3 subsets were chosen. This is a way of overfitting and should be visible in the error of the RMSE. However, the effect is not very large.

Cloud type

The division by cloud type is done by dividing the data set in 3 subsets: clear sky, water clouds and ice clouds. For each subset we compute the RMSE as described in section 4.2.1 including only data of that subset. These RMSE are then weighted averaged by their number of points (if a subset contains more points it gets a higher weight). The resulting RMSE is compared to the reference RMSE. An improvement is expected due to better prediction of clear sky values. Thus yielding a lower overall RMSE. The RMSE difference for each training/testing station combination is shown in figure 4.3b.

Solar zenith angle

For division by solar zenith angle, the data set is split for solar zenith angles in the following ranges:

$$\begin{array}{rll} & SZA & < 30^\circ \\ 30^\circ & < SZA & < 40^\circ \\ 40^\circ & < SZA & < 50^\circ \\ 50^\circ & < SZA & < 60^\circ \\ 60^\circ & < SZA & < 70^\circ \\ 70^\circ & < SZA & \end{array}$$

The solar zenith angle is expected to have an effect on prediction of the UVDr as the spectrum of solar irradiance at the surface changes with the solar zenith angle. As the solar zenith angle increases, more light is scattered and absorbed especially at shorter wavelengths. The UVDr is expected to be more dependent on band 2 for large solar zenith angles (i.e. the band 2-parameter would be larger for large SZA).

The difference in RMSE resulting from dividing data by solar zenith angle (with respect to the reference RMSE) is shown in figure 4.3c.

Direct/diffuse ratio

The ratio of direct/diffuse light at the surface may be expected to be more closely related to the solar spectrum at the surface than for example SZA or cloud optical thickness. However, the direct/diffuse light has a larger error than e.g. the solar zenith angle (which is exactly known). The amount of direct light is quantified with respect to the total irradiance:

$$f_{dir} = \frac{I_{dir}}{I_{dir} + I_{diff}}$$

The data set is divided by f_{dir} values within the following ranges:

$$
\begin{array}{rl}
 & f_{dir} < 0.1 \\
0.1 < & f_{dir} < 0.3 \\
0.3 < & f_{dir} < 0.5 \\
0.5 < & f_{dir} < 0.7 \\
0.7 < & f_{dir}
\end{array}
$$

The resulting RMSE with respect to the reference RMSE is shown in figure 4.3d.

COT and the choice of division

Cloud optical thickness is expected to affect the spectral composition of light as it largely determines the amount of Mie scattering versus the amount of Rayleigh scattering. The data is split by cloud optical thickness value within the following ranges:

$$
\begin{array}{rl}
 & COT = 0 \\
0 < & COT < 2 \\
2 < & COT < 5 \\
5 < & COT < 10 \\
10 < & COT < 20 \\
20 < & COT
\end{array}
$$

This choice of ranges is done manually by looking at a histogram of values and making a guess of a 'fair' division. The resulting RMSE minus the reference RMSE is shown in figure 4.4a.

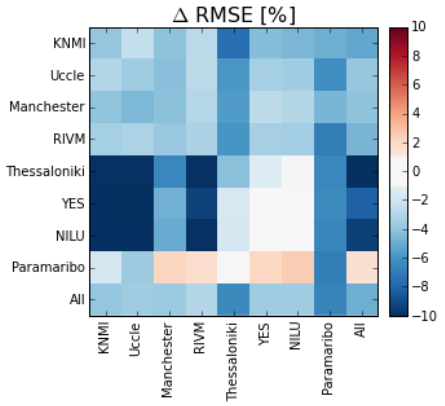
We can also compute a fair division by making all bins contain approximately the same amount of points. Because there is a vast amount of 0 COT points (clear sky), we will make this a separate bin. The resulting division is as follows:

$$
\begin{array}{rl}
 & COT = 0 \\
0 < & COT < 1.23 \\
1.23 < & COT < 2.81 \\
2.81 < & COT < 6.42 \\
6.42 < & COT < 18.7 \\
18.7 < & COT
\end{array}
$$

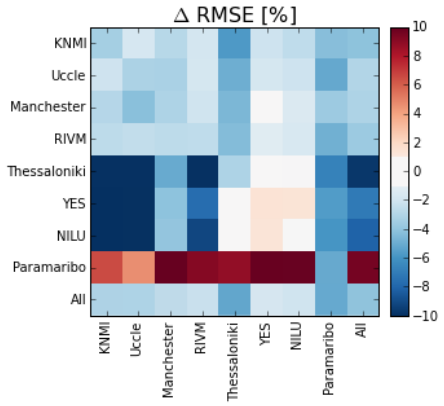
The resulting RMSE difference is shown in figure 4.4b.

From the figures (4.4a and 4.4b) we can see a small difference for most stations and a particularly large difference for models trained on Paramaribo data. The difference between these figures is shown in figure 4.4c, where values up to +10% are present.

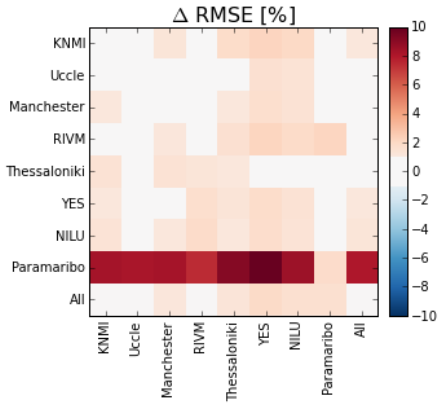
From this small difference at most stations we can conclude that the choice of division does not matter very much, but may improve or reduce the RMSE somewhat. Data from Paramaribo shows an unexpectedly large difference in RMSE between different choices of division. This is likely due to the large amount of outliers in the data. Because of its unexpected behavior, the Paramaribo data is not included in the training data of the final model, which is



(a) Division by cloud optical thickness using manually chosen bins.



(b) Division by cloud optical thickness using evenly distributed bins.



(c) Difference between both COT divisions.

Figure 4.4: The RMSE from division by cloud optical thickness minus the reference RMSE (top figures). The bottom figure shows the difference between both top figures.

unfortunate because of its unique location close to the equator and near the edge of the viewing disc.

Conclusions on the use of subsets

First, it is important to note that the qualification of the fits by the RMSE is questionable. The RMSE represents the residuals; a quantity that is minimized by OLS (ordinary least squares). However, the model is not trained using OLS, but using ODR (orthogonal distance regression), so an improvement of the fit (smaller orthogonal distances) does not necessarily lead to lower RMSE. We

could even argue that we should use fitting methods that are more robust to outliers like RANSAC, which would mean that each training/testing fit is to be quantified by eye. However, ODR is somewhat similar to OLS, so RMSE should still be a fine indicator of quality.

For training models on Paramaribo data, dividing by subsets does not yield better results when tested on other stations' data. Paramaribo has the smallest data set (less than 400 points) and the largest spread. The large spread is also seen from the reference RMSE, where Paramaribo used as testing station always shows large RMSE. The smaller subsets rather obscure the general behavior, steering the fit in an erratic direction. This fits the Paramaribo data itself slightly better, but is not representative for the behavior at other stations (behavior meaning, the response of the UVDr w.r.t. SICCS bands 1 and 2). The 'All' station does not include Paramaribo data for this reason (its erroneous data).

Division by solar zenith angle yields the largest RMSE reductions for most testing/training station combinations. However, training stations on the Thessaloniki data and testing it for KNMI or RIVM data yields a worse result (higher RMSE). Also, training the model on any station's data other than KNMI and Thessaloniki data (including YES and NILU) yields a larger RMSE for 'all' data.

The ideal overall division seems to be the cloud type division or the COT division, which are similar. Also, the Paramaribo station is likely to be excluded from the eventual model fit because of its chaotic behavior and/or large amount of outliers.

4.2.2 Different model

The ill reproducibility of one stations data from another stations model is likely a result of overfitting, related to the multicollinearity of the explaining parameters (band 1 and 2 are strong correlated). A solution for this ill reproducibility is to use a different model. In this section we study different model equations. We only discuss models that include only band 1. Irradiance in the wavelength range of band 2 only account for roughly 20 percent of the total UVDr, and includes some longer wavelengths that are not important for the UVDr at all. Band 1 is therefore expected to better relate to the UVDr. The following equations are discussed:

$$I_{UVDr} = C_1 + C_2 \cdot I_1 + C_3 \cdot I_1^2 \tag{4.2}$$

$$I_{UVDr} = C_1 + C_2 \cdot \exp\left(C_3 \cdot I_1\right) \tag{4.3}$$

$$I_{UVDr} = C_1 + C_2 \cdot I_1^{C_3} \tag{4.4}$$

The model equation we try first is equation 4.2. The success of the new model equation is evaluated by computing the RMSE as is done in the previous section (section 4.2.1). The RMSE for testing and training the model for the same station's data is expected to increase when using a simpler model. This

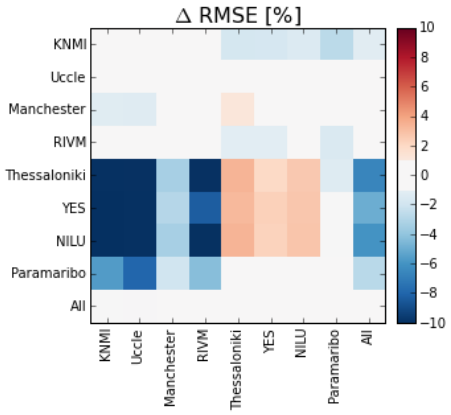


Figure 4.5: The RMSE for each training/testing station combination using model equation 4.2 minus that using model equation 4.1.

is simply because less data is used and overfitting is not permitted anymore. Testing the model with data other than is used to train it is expected to reduce the RMSE for the same reason.

The resulting RMSEs are shown in figure 4.5. We can see only minor differences in RMSE as most squares are white, i.e. within $\pm 1\%$ difference. On the diagonal there are only positive values as expected (red or white squares, where white squares are indeed positive values which is not visible in the figure). Major improvements are made for models trained on Thessaloniki data (all 3 kinds) and tested on (most) other stations (Paramaribo is unaffected likely because it already contained many outliers, regardless of model).

These improvements seem to be due to the amount of outliers in each data set (concluded from visual comparison of different station's data). The Thessaloniki data generally agrees well with ground observations except for a number of (severe) outliers (e.g. when comparing ground-based band 1 with SICCS band 1). Most northern stations agree worse (larger spread around the 1:1 line) but contain fewer/less severe outliers. A more complex model has more opportunity to compensate for such outliers, resulting in overfitting for these stations. An alternative to choosing a different model equation, we may choose a different fitting method that is more robust to outliers. This is done in a following section.

Concluding, we can state that model equation 4.2 yields a better result overall than equation 4.1, using the current fitting method (orthogonal distance regression). The initial problem, that is, the large variation in model parameters between stations, is mostly solved using model equation 4.2. We can see this in figure 4.6.

The take-home message is not to make the model too complex to prevent overfitting. The preferred model equation is one that reflects the physical relation between the UVDr and the separate band(s). However, the physical relation between band 1 and the UVDr is that of adding an (arbitrary) spectrum inside

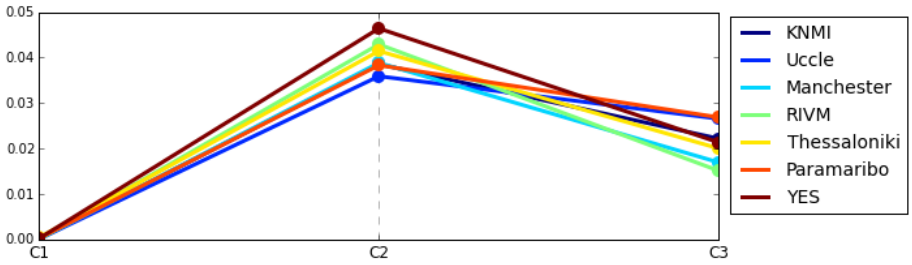


Figure 4.6: The 3 model parameters resulting from training the model of equation 4.2 using data of single stations. Each parameter is multiplied by the mean of all data of the respective insolation or product of insolations such that parameters of large quantities are not dwarfed.

the integral (the integral of ground spectrum over wavelengths). This is not simply expressed by a functional relation, and therefore any equation that *looks* good can be argued to *be* good.

Equations 4.3 and 4.4 were not presented here but have been studied and show similar consistency. The question is now which of these equations 'looks' best. My preference goes to equation 4.3 or 4.4 because the relation between UVDr and band 1 is not necessarily exactly quadratic. Equation 4.4 appears to be more difficult to solve numerically (using python's fitting methods) so equation 4.3 is preferred.

4.2.3 Fitting methods

Different models can be found using different fitting methods like OLS, ODR and RANSAC. These methods are compared here. The way to check the quality of the method is not done by considering the RMSE as this will always favor OLS. Instead we look at variation in fitting parameters resulting from fits on each station's data. The quality of each method is then determined by the consistency of the model between different stations. This consistency indicates how well the model can predict values outside its fitted range.

The usage of methods that are robust to outliers (e.g. RANSAC, Theil-Sen) have an advantage over conventional methods (OLS, ODR) when comparing satellite data to ground-based data. Satellite data may have actual outliers; values that deviate from the trend for no particularly physical reason. Outliers are less likely to occur when comparing ground-based data with other ground-based data (e.g. ground-based band 1 with UVDr). Outliers may still occur in this case, but are more likely to have a physical reason for their deviation from the trend.

RANSAC

The RANSAC (random sampling consensus) is a fitting method that draws random fits through the data and picks the best result at the end of the iteration. The RANSAC method randomly picks a number of points (equal to the number of fitting parameters - in our case - so a perfect fit can be made). All points within a chosen distance from this fit (the inliers) are counted, and if the total count exceeds a certain number (percentage of total number of points), it is accepted. This process is then repeated a number of times. The fit parameters that yielded the highest number of inliers are ultimately chosen.

Multiple issues of a RANSAC fit on a certain data set will seldom yield the exact same fit parameters. The variation of fit parameters for different issues are too large for my liking. My variation on the RANSAC method does not pick the best result, but instead (weighted) averages the results of all iterations where the criterion was passed (the RANSAW method; random sample weighting). The averaging is done perfectly for polynomial fit functions by taking the average of the fit parameters. The average of non-polynomial functions does not yield a function of the same form. The average is instead approximated by dividing the fitted data set range into 101 points for each fit that passed the criterion. Each of these 101 points applied to the fit function is then (weighted) averaged. A new fit is then constructed from these 101 points using ODR which yields the final fit parameters.

4.2.4 Fit to ground measurements

We can fit the ground-based UVDr to the ground-based band 1 (and 2) measurements, rather than to the SICCS band 1 and 2 output. There is a physical relation between band 1 and the UVDr, but this relation is heavily obscured by the satellite approximation. This relation is not obscured when comparing ground-based UVDr with ground-based band 1.

4.3 Combinations of solutions

One method is found that is able to predict values across stations is to combine the following aspects discussed in the previous sections:

- Use the clear sky subset
- Use the RANSAC fitting method
- Use model equation 4.3
- Optional: Use ground measurements

Most (all) stations' response of UVDr w.r.t. band 1 is well described by the clear sky signal. Data for which clouds are present generally follow the same signal but with more spread/noise. Using ground measurements shows good agreement between stations, but using SEVIRI measurements can also result in

	C_1	C_2	C_3	D	%
SICCS band 1	-0.130 ±0.41%	0.131 ±0.40%	0.161 ±0.25%	0.005 W/m^2	65%
G.-b. band 1	-0.110 ±0.22%	0.108 ±0.21%	0.171 ±0.14%	0.004 W/m^2	75%

Table 4.2: Fitting parameters resulting from the RANSAW method of step 1 (equation 4.3). The distance criterion and the inlier ratio are given by the last two columns (D and %).

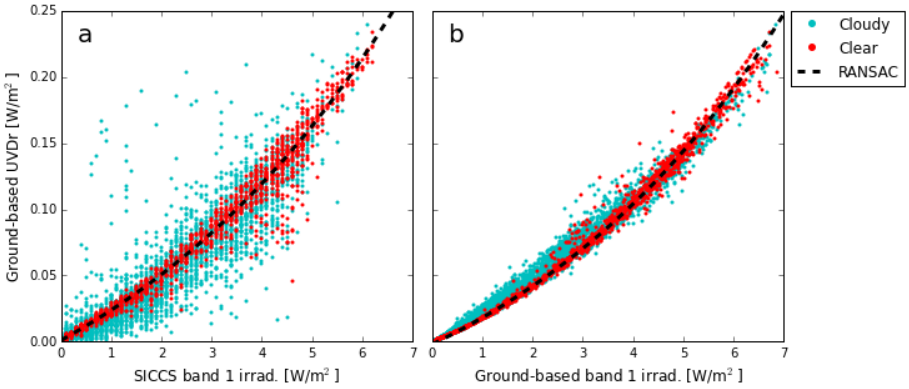


Figure 4.7: The UVDr as a function of band 1, using SICCS band 1 (panel a) and ground-based band 1 (panel b). Clear sky points are shown in red, cloudy sky in light blue and the (clear sky) fit is indicated by the black dashed line.

good agreement using the proper fitting method. The RANSAC method (or my adjusted 'RANSAW' method) can be tuned (by changing outlier distance and fraction) to pick out the dominating signal. This is step 1 of the fitting method. This dominant signal generally follows the same curve across stations.

This method would only account for spectral effects resulting from different SZA, such as Rayleigh scattering. The presence of clouds affects the amount of scattering and as such, the clear sky relation may not hold for cloudy sky. To correct for the spectral effect of clouds is step 2 of the method.

4.3.1 Step 1

Figure 4.7 shows step 1 of the fitting method; to fit clear sky UVDr to clear sky band 1 irradiance. Fits to ground based and satellite based band 1 observations are shown. The fit parameters (mean of the RANSAW method ± standard deviation of mean for each parameter, percentage indicates the percentage of the parameter in question) for SICCS band 1 and ground-based band 1 are shown in table 4.2.

The resulting fit is assumed to only show the spectral effect of the solar zenith angle. There are more possible spectral effects (e.g. absorption by ozone and aerosols), but we decide to ignore these effects as they are obviously small:

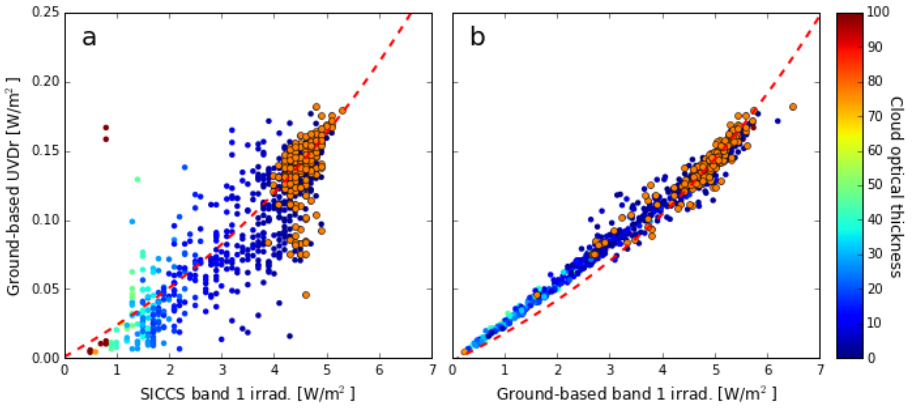


Figure 4.8: Data with solar zenith angles between 30 and 35 degrees with SICCS based band 1 (panel a) and ground-based band 1 (panel b). Clear sky points are shown in orange, all other points are colored according to their cloud optical thickness. The fit of step 1 is shown by the red dashed line.

such is the spread in the clear sky scatter plot. The spectral effect that causes the resulting non-linear relation can be well explained by Rayleigh scattering which is mostly dependent on SZA (and atmospheric composition, which hardly varies). The spread introduced by the presence of clouds can also be accounted for, which is step 2 of the fitting procedure.

4.3.2 Step 2

To include the presence of clouds, we assume the cloudy irradiance to be equal to the clear sky irradiance multiplied by the 'irradiance factor' f_x:

$$I_x(\tau) = I_x(0) \cdot f_x(\tau) \tag{4.5}$$

where subscript x indicates a specific irradiance (e.g. UVDr or SICCS band 1). The resulting model equation, i.e. the cloud-corrected version of equation 4.3, becomes:

$$I_{UVDr}(\tau) = \left(C_1 + C_2 \cdot \exp\left(C_3 \cdot \frac{I_1(\tau)}{f_1(\tau)}\right)\right) \cdot f_{UVDr}(\tau) \tag{4.6}$$

We want to relate these irradiance factors to SICCS input and/or output so they are available for every SICCS calculation. We attempt to derive the irradiance factor from cloud optical thickness and solar zenith angle.

Cloud optical thickness

In figure 4.8 we plotted irradiances that occurred within a small range of solar zenith angles. The clear sky points lie clumped together whereas cloudy points

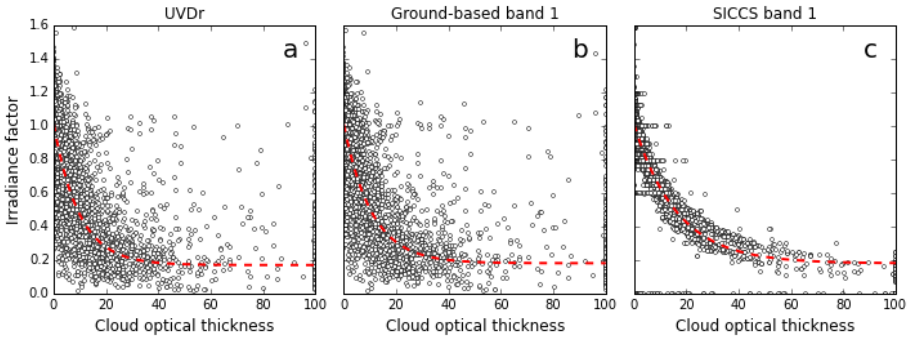


Figure 4.9: The irradiance factor (see text) as a function of cloud optical thickness for ground-based UVDr (a), ground-based band 1 (b) and SICCS band 1 (c). The red line indicates the RANSAW fit.

deviate from this clump towards zero. The cloud optical thickness (shown in color) is higher for points closer to the origin (i.e. have a smaller irradiance factor). To study the relation between the irradiance factor and the cloud optical thickness, we need a method to compute irradiance factors. This is done as follows (equation 4.7):

$$f_{x,i} = I_{x,i} \ / \ \overline{I_{x,cs}}(\phi_i) \tag{4.7}$$

Which computes an irradiance factor f_x for each point i, where ϕ_i is the solar zenith angle, $I_{x,i}$ is the actual irradiance (of type x, e.g. UVDr or SICCS band 1) for that point and $\overline{I_{x,cs}}(\phi_i)$ represents the would-be clear sky value (the overbar indicates the mean of the bin that includes $SZA = \phi_i$). The would-be clear sky value $\overline{I_{x,cs}}(\phi_i)$ is computed by taking the mean irradiance of a clump of points from a small range of solar zenith angles. All other (cloudy) points will be divided by the would-be clear sky value according to the data point's solar zenith angle. For clarity: every point has a specific solar zenith angle, say 42.3. This point belongs to the wavelength range 42-43, and is thus divided by the mean of all clear sky points in the same solar zenith angle range to yield that point's irradiance factor. The irradiance factor for every type of irradiance (UVDr, ground-based band 1, SICCS band 1) is plotted against the COT in figure 4.9.

The fits are constructed using the RANSAW method. The irradiance factor is fitted as a function of COT with the following equation:

$$f_x = C_1 + C_2 \cdot \exp(-C_3 \cdot \tau) \tag{4.8}$$

We found the best results when fitting data with water and ice clouds separately. The parameters resulted from 100,000 iterations and the criteria specified in table 4.3

		C_1	C_2	C_3	D	%
Water	UVDr	0.207 ±0.93%	0.848 ±0.23%	0.078 ±0.86%	0.2	70%
	G.-b. band 1	0.214 ±0.89%	0.845 ±0.23%	0.076 ±0.81%	0.2	70%
	SICCS band 1	0.226 ±0.25%	0.819 ±0.06%	0.059 ±0.20%	0.1	85%
Ice	UVDr	0.135 ±0.90%	0.890 ±0.17%	0.135 ±0.64%	0.15	60%
	G.-b. band 1	0.151 ±0.95%	0.871 ±0.19%	0.126 ±0.71%	0.15	60%
	SICCS band 1	0.171 ±0.39%	0.871 ±0.07%	0.078 ±0.22%	0.1	85%

Table 4.3: Fit parameters and the RANSAW criteria used for the COT vs. irradiance factor fitting procedure (equation 4.8).

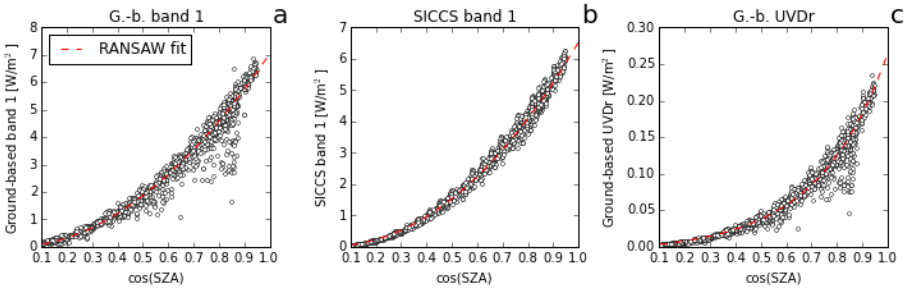


Figure 4.10: Fitting clear sky irradiances vs cosine solar zenith angle. The RANSAW fits are indicated with red dotted lines. Note that the UVDr fit (c) uses a different fitting equation (equation 4.12 instead of 4.10).

Solar zenith angle

Another way of correcting for cloud cover can be done using solar zenith angle data. There is a strong relation between the solar zenith angle and any irradiance quantity as it depicts an upper boundary for the irradiance value. This upper boundary can be approximated by fitting clear sky irradiance as a function of cosine SZA (figure 4.10). The irradiance factor f_x that was introduced in the previous section is then:

$$f_{x,i} = \frac{I_{1,i}}{I_1(0, SZA_i)}$$

for data point i, where $I_1(0, SZA)$ (i.e. the clear sky ($\tau = 0$) band 1 irradiance at the specified solar zenith angle) is a yet unknown quantity. We further assume f_{UVDr} and f_1 to be equal, so that f_{UVDr} can be computed from SICCS band 1 data. The resulting model equation with SZA correction becomes:

$$I_{UVDr}(\tau) = \left(C_1 + C_2 \cdot \exp\left(C_3 \cdot \frac{I_1(\tau)}{I_1(\tau)/I_1(0, SZA)}\right)\right) \cdot \frac{I_1(\tau)}{I_1(0, SZA)}$$

$$= \left(C_1 + C_2 \cdot \exp\left(C_3 \cdot I_1(0, SZA)\right)\right) \cdot \frac{I_1(\tau)}{I_1(0, SZA)} \tag{4.9}$$

	C_1	C_2	D	%
G.-b. band 1	7.037 ±0.02%	1.916 ±0.03%	0.2 W/m^2	65%
SICCS band 1	6.453 ±0.02%	2.085 ±0.05%	0.2 W/m^2	80%

Table 4.4: Fit parameters and RANSAC criteria for fitting clear sky irradiance against cosine SZA (equation 4.10).

Here, $I_{UVDr}(\tau)$ on the l.h.s. is explicitly written as a function of cloud optical thickness (τ) to indicate that the relation holds for every I_{UVDr} and not just for clear sky.

The parameter $I_1(0, SZA)$ is found by fitting band 1 irradiance to the solar zenith angle, and can be done for both the ground-based band 1 data and SICCS band 1 data. The equation to be fitted is:

$$I_1(0, SZA) = C_1 \cdot \cos(SZA)^{C_2} \tag{4.10}$$

The RANSAW method is used, performing 100,000 iterations. The result is shown in table 4.4.

4.3.3 Evaluating cloud corrections

The clear sky fit of step 1 is already expected to give a reasonable result. The addition of the cloud correction should improve the result. We check this by comparing both methods. The UVDr computed using the correction based on the COT is shown in figure 4.11 and is compared with UVDr computed without this correction. The same comparison is done for the correction based on SZA in figure 4.12.

The COT correction seems to slightly increase the correlation of the satellite data, although it is very small. The ground-based data has an increased number of outliers due to this correction and thus has a very low correlation coefficient. Such outliers occur when a very large COT occurs, which causes the corrected band 1 value to blow up. This in turn causes a (much) higher UVDr value than expected. The reason this does not happen for the satellite data, is most likely because the satellite band 1 value is directly computed from the cloud optical thickness in the SICCS algorithm.

The SZA correction slightly decreases the correlation of satellite data with UVDr data, but increases the correlation with the ground-based data. Although the correlation is already close to one, the difference from 1 is ∼halved. The uncorrected model also shows a bias for lower (about <0.15 W/m^2) values which has disappeared in the corrected model. Note that all points have been corrected by this method, even clear sky points. Giving clear sky points a factor of 1 results in some outliers for the ground-based data and a slightly reduced correlation for the satellite data.

There are 2 ways to proceed from hereon: (i) Use the COT-corrected model, taking satellite data as input, or (ii) use the SZA-corrected ground-based model, which takes satellite data that is fit to represent ground-based data. The first (i)

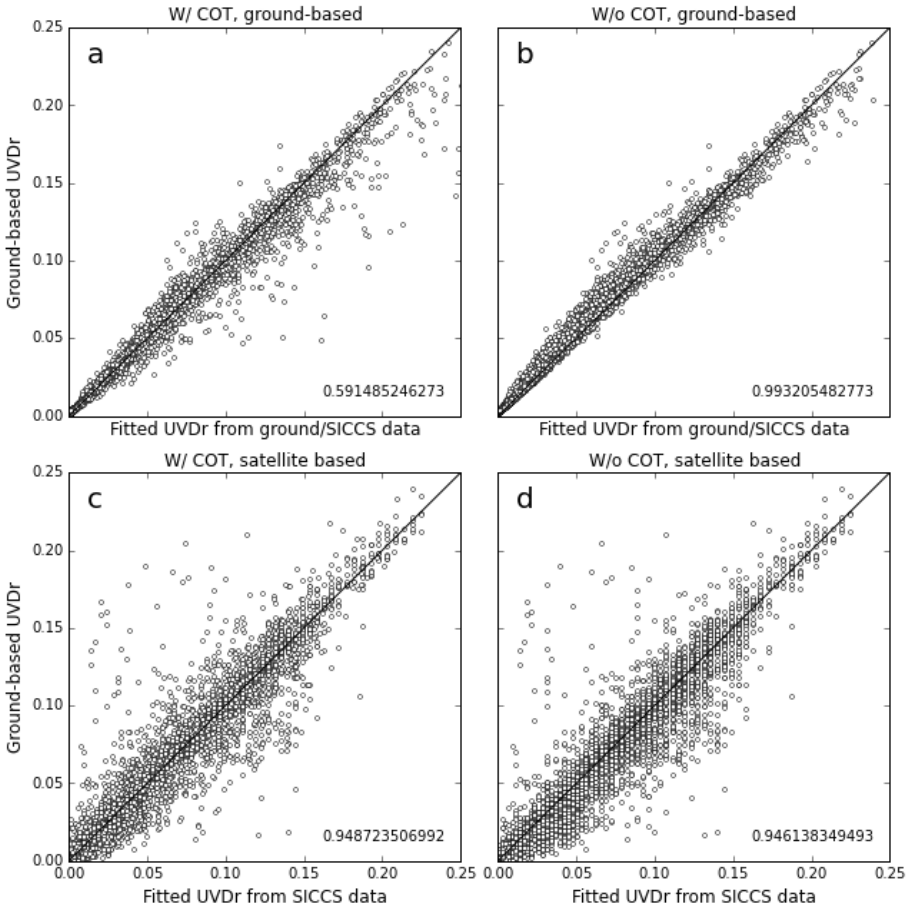


Figure 4.11: Comparison between modelled and measured UVDr using different methods. Top figures (a+b) include the ground-based band 1 values whereas the bottom figures (c+d) use only SICCS data. The left figures (a+c) are with the COT correction and the right figures are computed using only the fit of equation 4.3 including only clear sky data. Some points from panel (a) are not shown in the figure as they lie far outside the domain, but result in a low r^2.

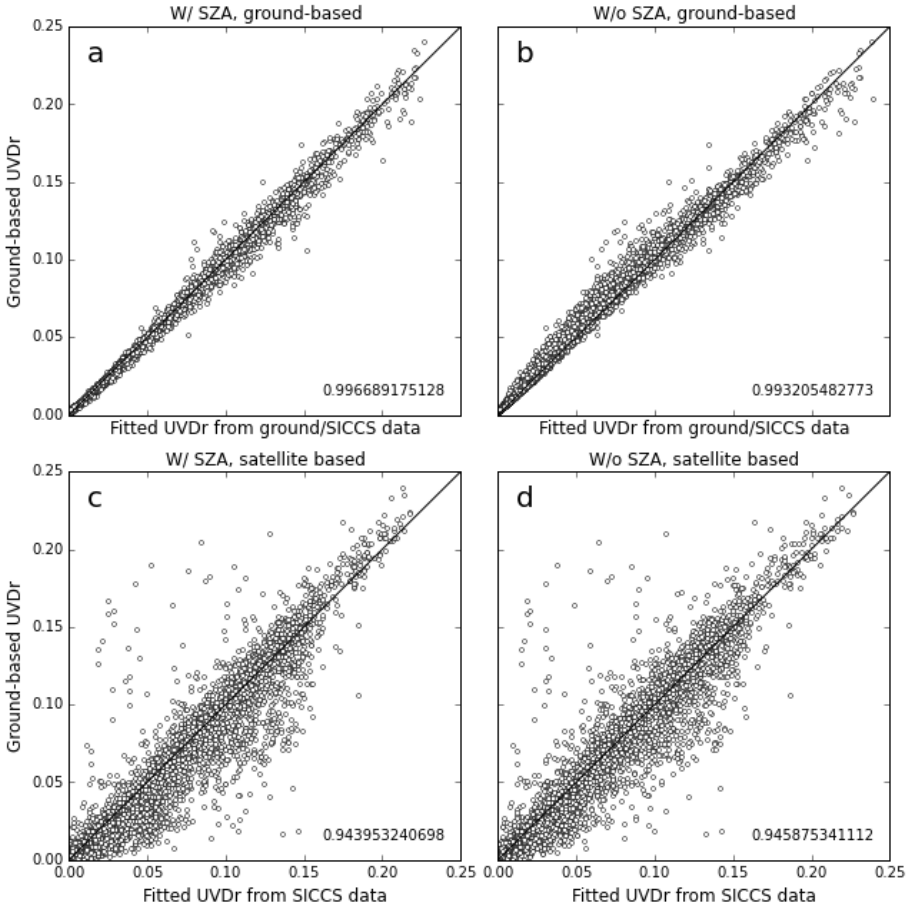


Figure 4.12: Comparison between modelled and measured UVDr using different methods. Top figures (a+b) include the ground-based band 1 values whereas the bottom figures (c+d) use only SICCS data. The left figures (a+c) are with the SZA correction and the right figures are computed using only the fit of equation 4.3 including only clear sky data.

method is expected to give better results for the fitted domain. The second (ii) method *may be* expected to give better results for stations/times that are not included in the fitted domain, because it better represents the actual physics. These methods are compared to method (o), which uses only step 1 and no cloud correction (i.e. solely equation 4.3 for both clear and cloudy sky points). Two more methods have come to mind that are also tested: (iii) use the SZA-corrected SICCS-based model. This method is shown to be worse on all data simultaneously (figure 4.12, the r^2 is lower in panel c than in panel d), but may prove to be robust when trained/tested on various stations. And (iv), a method that is not discussed, but is based on the SZA correction philosophy and is explained best by the following equation:

$$I_{UVDr}(\tau) = f_1(\tau) \cdot I_{UVDr}(0) = \frac{I_1(\tau)}{I_1(0)} \cdot I_{UVDr}(0) \tag{4.11}$$

This method completely skips step 1, which is designed to reproduce clear sky UVDr from clear sky band 1 values. For the SZA correction method we derive clear sky values from SZA only, so we may as well also do this for the UVDr. This is shown in figure 4.10c and uses a different fit equation than the band 1 irradiances, namely:

$$I_{UVDr}(0) = C_1 \cdot \left(e^{-C_2 \cdot \cos(SZA)} - 1\right) \tag{4.12}$$

We also use the assumption that the cloud effect is equal on the UVDr and band 1 irradiance, such that the $f_{UVDr}(\tau)$ (equation 4.5) can be approximated by:

$$f_{UVDr}(\tau) \approx f_1(\tau)$$

Each method (consisting of a multitude of fits) is trained and tested on various (combinations of) stations. The ability of a model trained on one station to reproduce data of another station is quantified by computing the Pearson correlation coefficient, the Spearman correlation coefficient and the bias. The correlations are used to find if a good linear relation is found between the model output and the ground-observed data of the station. If strange values/outliers are produced by the model, the Pearson correlation drops whereas the Spearman correlation is not affected. Neither correlation can be used to find systematic wrong estimation, so we also regard the bias. For each method, a matrix containing each of these 3 quantities is computed, which is shown in color code in figure 4.13. The bias is shown as absolute values on a double color scale to better distinguish high and low values (not positive/negative). The sign of the bias is not important for qualifying the models. The bias is the average bias per point (bias divided by the total number of points) and multiplied by 1000 (so the result is in mW/m^2). This is solely for comparison purposes.

4.3.4 Conclusion

Method (ii), (iii), (iv) rarely perform better than method (o) regarding correlation and bias. Method (i) seems to train badly on some stations (e.g. Reading,

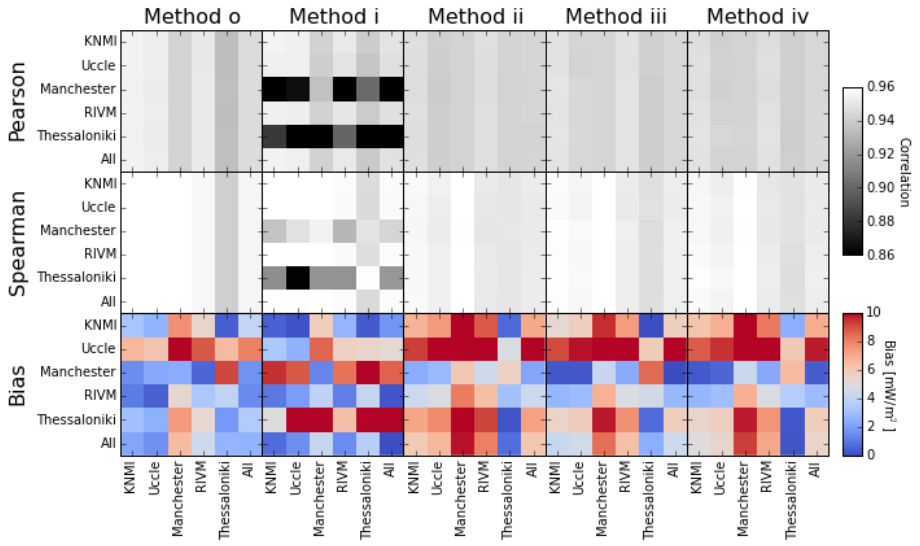


Figure 4.13: Pearson correlation, Spearman correlation and bias for each training/testing station combination for each method. For an explanation of the separate methods, see text.

Thessaloniki), resulting in low correlations and high biases. However, method (i) shows the best overall result when trained on all stations together. Upon inspecting data of Reading and Thessaloniki I find their relation between irradiance factor and COT is poorly expressed. Yet, including the COT correction based on other data does improve the prediction of data of these stations (compare method (o) and (i) trained on 'all', tested on e.g. Reading: method (i) is better). I will therefore use method (i), but train step 2 using only data from stations where this relation is well expressed (KNMI, Uccle, RIVM). I have followed similar logic to great success; the step 1 fit is also only based on data where the behavior is well expressed (i.e. clear sky data).

The COT fitting procedure is performed again, now including only data of the above mentioned stations. See table 4.5 for the result. We have in principle finalized the UVDr fitting procedure. The resulting model is repeated here:

$$I_{UVDr}(\tau) = \left(-0.130 + 0.131 \cdot \exp\left(0.161 \cdot \frac{I_1(\tau)}{f_1(\tau)}\right)\right) \cdot f_{UVDr}(\tau) \qquad (4.13)$$

with

$$f_1(\tau) = \begin{cases} 1 & \text{if } cph = 0 \\ 0.225 + 0.824 \cdot \exp(-0.057 \cdot \tau) & \text{if } cph = 1 \\ 0.182 + 0.882 \cdot \exp(-0.079 \cdot \tau) & \text{if } cph = 2 \end{cases} \qquad (4.14)$$

		C_1	C_2	C_3	D	%
Water	UVDr	0.178 ±0.87%	0.896 ±0.18%	0.072 ±0.70%	0.15	60%
	G.-b. band 1	0.199 ±0.74%	0.891 ±0.19%	0.076 ±0.67%	0.15	60%
	SICCS band 1	0.225 ±0.23%	0.824 ±0.05%	0.057 ±0.18%	0.1	85%
Ice	UVDr	0.138 ±0.80%	0.925 ±0.15%	0.133 ±0.57%	0.15	60%
	G.-b. band 1	0.155 ±0.78%	0.894 ±0.16%	0.127 ±0.62%	0.15	60%
	SICCS band 1	0.182 ±0.33%	0.882 ±0.06%	0.079 ±0.20%	0.1	85%

Table 4.5: Fitting parameters describing irradiance factors as a function of COT, according to equation 4.8

$$f_{UVDr}(\tau) = \begin{cases} 1 & \text{if } cph = 0 \\ 0.178 + 0.896 \cdot \exp(-0.072 \cdot \tau) & \text{if } cph = 1 \\ 0.138 + 0.925 \cdot \exp(-0.133 \cdot \tau) & \text{if } cph = 2 \end{cases} \tag{4.15}$$

where τ is the cloud optical thickness and I_1 the (all-sky) band 1 irradiance as given by the SICCS product.

4.3.5 Latest version based on additional available SICCS data

For step 2 of the fitting procedure I have introduced the irradiance factor f_x. This was computed from the clear sky value and actual sky value. The clear sky value was derived from either an average of irradiances with the same SZA, or by fitting irradiance as a function of SZA using only clear sky points.

The SICCS algorithm also computes clear sky values at every pixel at any time, which were not available to me during every step of the fitting procedure. These clear sky irradiances will be available soon, which will provide me with a new method. The COT method that is chosen uses the COT to find the irradiance factor f_x. With the clear sky values I can compute the irradiance factor directly, and as such a more accurate model is expected. This method will be added in this section.

Repeat of step 1 is in table 4.6. Only the fit of ground-based UVDr as a function of SICCS band 1 had to be repeated. However, for fair comparison with the other methods, all fits will be repeated because the SICCS algorithm in general was slightly altered.

With all fits repeated for the slightly altered data (that includes clear sky data sets, and has a slightly different clear sky calculation) we repeat the comparison between different methods. The change on method (i) - replacing the clear sky value fitted from COT with the clear sky value from the added data set - is added as a new method (i.2). The Pearson and Spearman correlations and the biases for each training/testing station combination for each method is shown in figure 4.14. Scatter plots of modeled UVDr vs. ground-observed UVDr for each method's best guess version (trained on all data) are shown in figure 4.15. These scatter plots are shown to give the reader a feeling for the

	C_1	C_2	C_3	D	%
SICCS band 1	-0.135±0.42%	0.136±0.40%	0.157±0.26%	0.005W/m^2	65%
G.-b. band 1	-0.110±0.22%	0.108±0.21%	0.171±0.14%	0.004W/m^2	75%

Table 4.6: Repeat of step 1 parameters (equation 4.3).

		C_1	C_2	C_3	D	%
Water	UVDr	0.199±0.72%	0.873±0.18%	0.077±0.66%	0.2	70%
	G.-b. band 1	0.212±0.67%	0.877±0.18%	0.079±0.65%	0.2	70%
	SICCS band 1	0.234±0.20%	0.826±0.05%	0.057±0.17%	0.1	85%
Ice	UVDr	0.137±0.83%	0.925±0.15%	0.133±0.58%	0.15	60%
	G.-b. band 1	0.156±0.78%	0.892±0.16%	0.128±0.62%	0.15	60%
	SICCS band 1	0.213±0.23%	0.860±0.05%	0.085±0.17%	0.1	85%

Table 4.7: Repeat of step 2: COT-method fit parameters (equation 4.8).

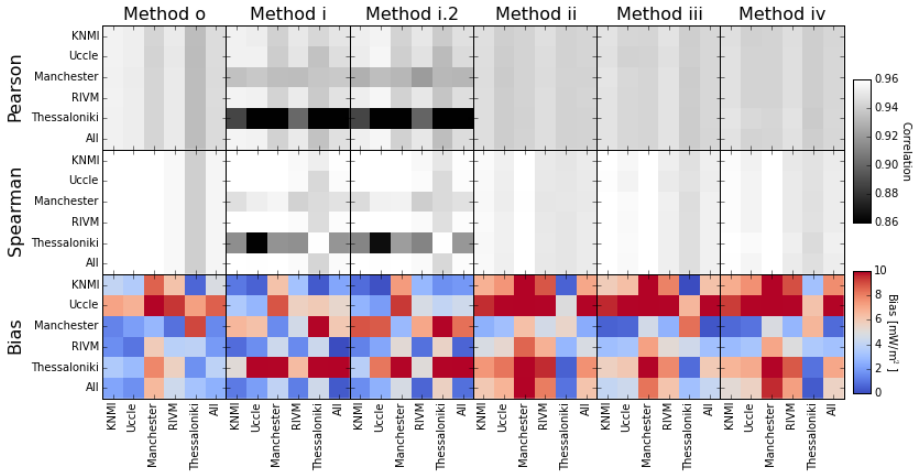


Figure 4.14: Repition of the comparison between methods that is done in figure 4.13.

validity of each method and the quality of our final choice (i.e. method [i.2]).

We can see that method (i.2) performs slightly worse than method (i), but still slightly better than (o). We choose method (i.2) over method (i) because it is assumedly more robust (it uses one fit fewer) even though the result on the present data is slightly worse.

For completeness I will write out the final model once more:

$$I_{UVDr}(\tau) = (-0.135 + 0.136 \cdot \exp(0.157 \cdot I_{1,cs})) \cdot f_{UVDr}(\tau) \tag{4.16}$$

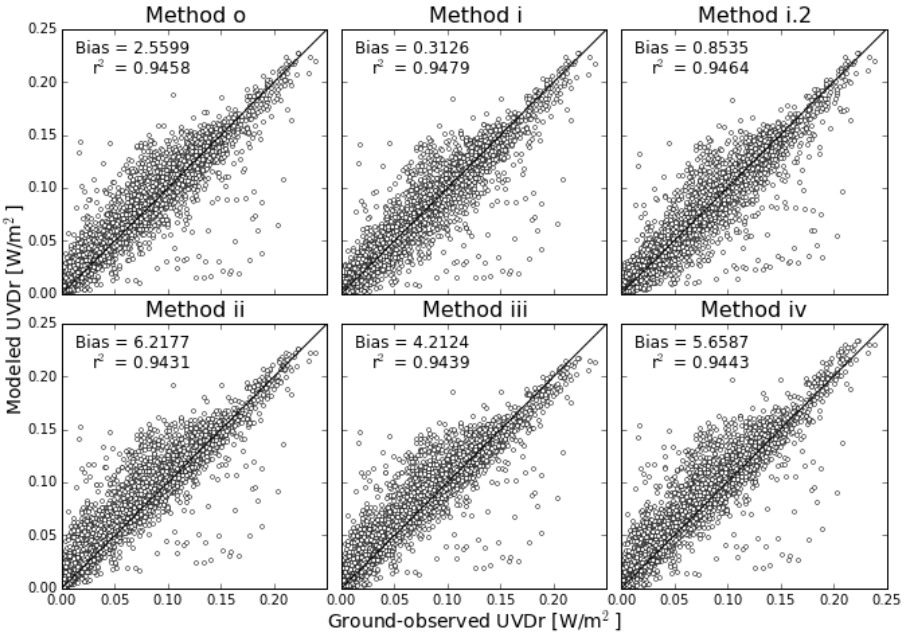


Figure 4.15: Scatterplots of modeled UVDr of each method when trained on all stations' data, versus the ground-observed UVDr.

with

$$f_{UVDr}(\tau) = \begin{cases} 1 & \text{if } cph = 0 \\ 0.199 + 0.873 \cdot \exp(-0.077 \cdot \tau) & \text{if } cph = 1 \\ 0.137 + 0.925 \cdot \exp(-0.133 \cdot \tau) & \text{if } cph = 2 \end{cases} \tag{4.17}$$

4.4 UVDr in January

The UVDr model stated in section 4.3.5 is trained solely on July data. Although data of January is also available, its data is generally much lower in value. We expect to train the most reliable model on the data set that contains the largest range of values. We use the January data here as another test of the quality of the model. Scatter plots of the modeled insolation versus the ground-observed insolation are shown in figure 4.16.

We can see that most of the ground-based UVDr data are well predicted by the UVDr model. Because the model is trained on SICCS input/output, the bias that is present for band 1 (figure 3.4) has been accounted for.

The interested reader may want to see the same figure for July (which has been visualized on many occasions, but not in the same way as we presented January), this is shown in figure 4.17

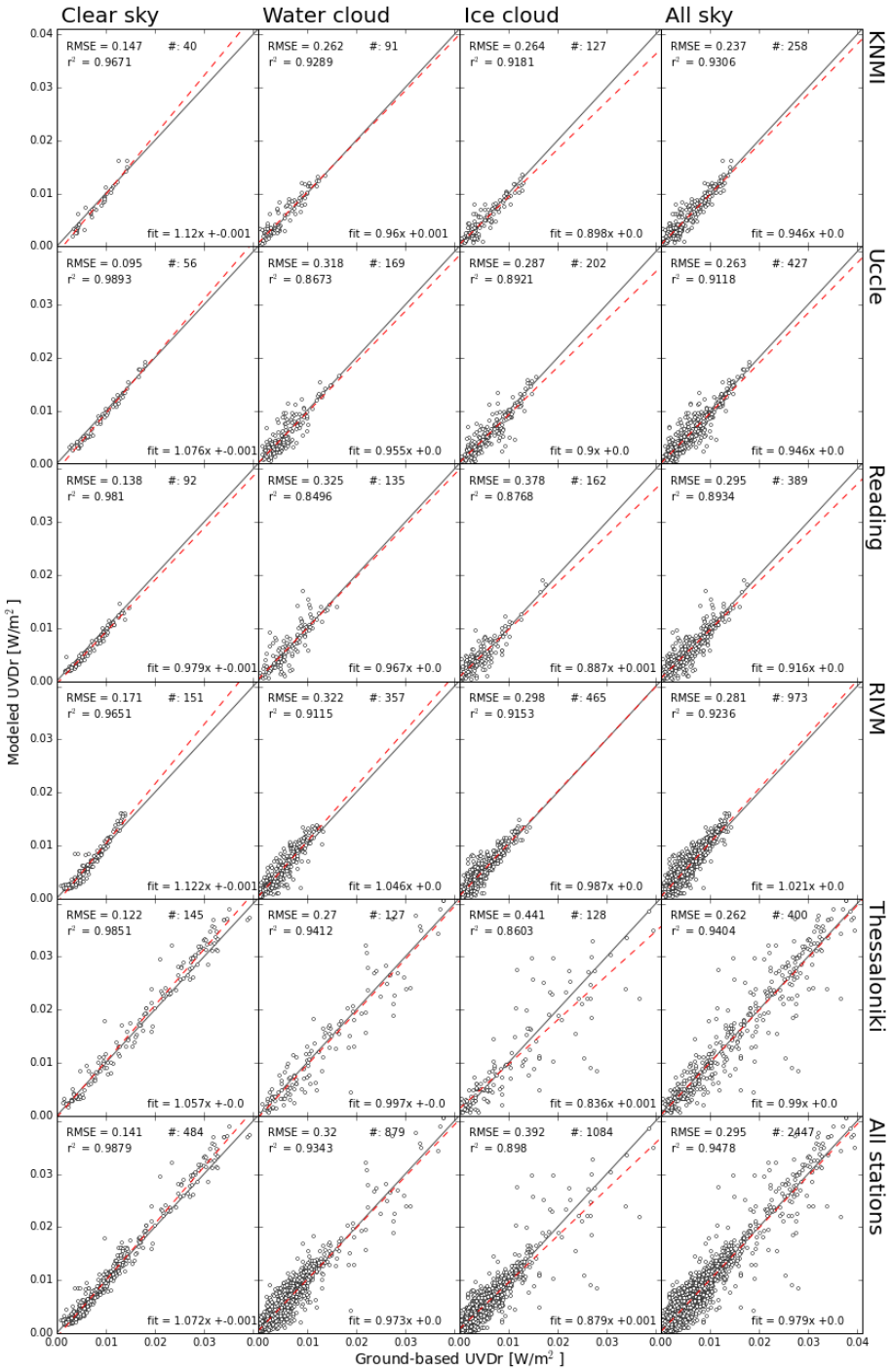


Figure 4.16: Modeled UVDr versus ground-observed UVDr during January 2014.

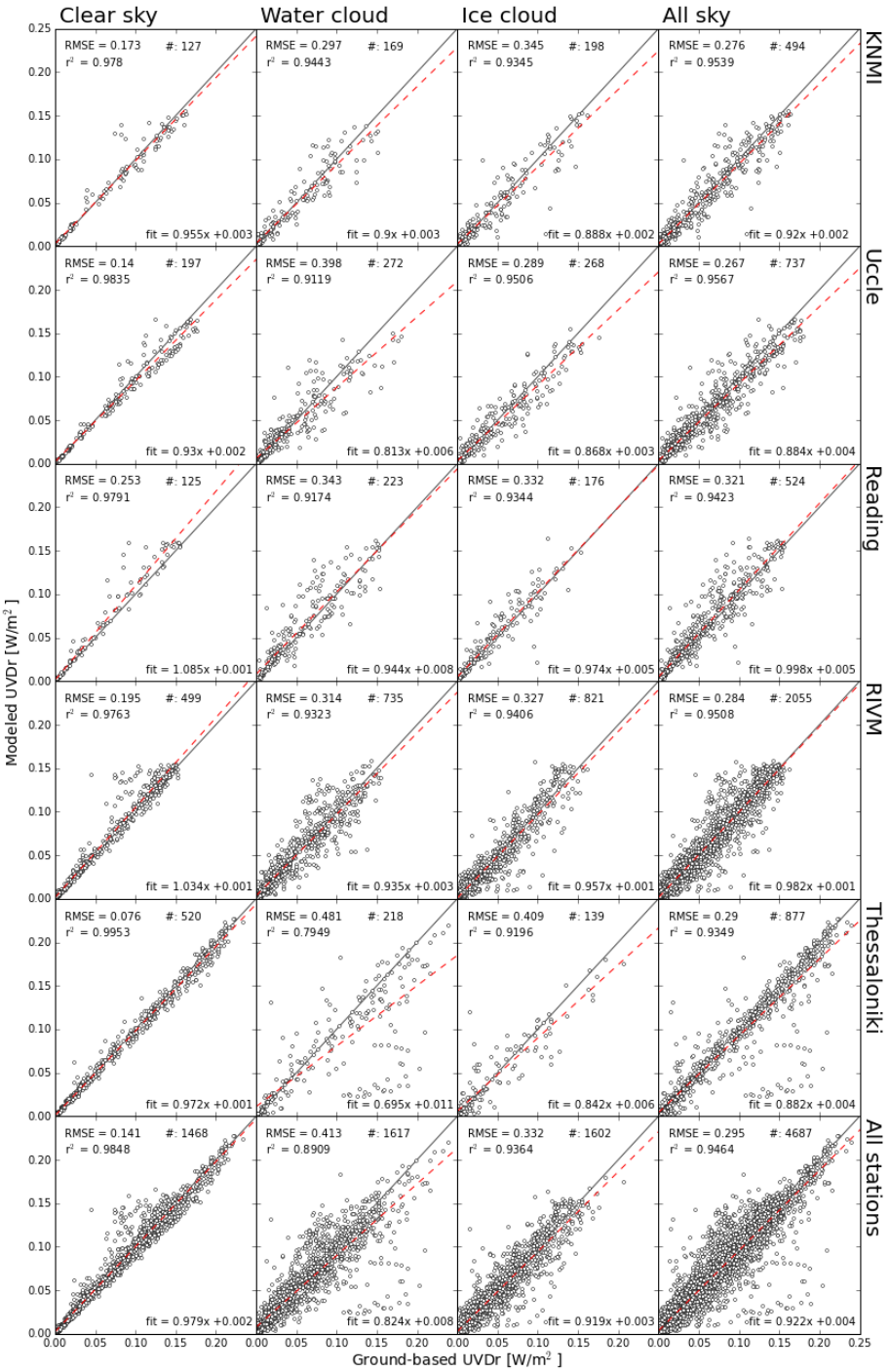


Figure 4.17: Modeled UVDr versus ground-observed UVDr during July 2014.

Chapter 5

Conclusions

Comparison between SICCS band 1 and ground-observed band 1 irradiance has shown that the SICCS algorithm generally underestimates the irradiance. This underestimation (bias) becomes larger for larger irradiance values, ergo it is a relative underestimation rather than an absolute underestimation. The contribution to the bias of individual parameters has been tested and no parameters have been found to cause a systematic bias, except for the water vapour path. The bias would generally become slightly more positive for higher water vapour paths. We have not studied the cause of this.

The UVDr can be deduced from SICCS output and cloud physical properties with good confidence that it is reproducible outside the domain (spatially and temporally) of the fitting data. A fit has been made that only includes band 1 data (i.e. excludes band 2) to avoid overfitting. The fit is performed on clear sky points only which describes the relation between the UVDr and band 1 best. A correction for cloudy sky points is included which slightly improves the correlation of all-sky SICCS irradiance with all-sky ground-based irradiance.

Acknowledgements

We would like to acknowledge several scientists for their cooperation of sharing their ground-based irradiance data. We acknowledge Ankie Piters and Marc Allaart for performing the Brewer measurements at KNMI and for providing us with Brewer data from Paramaribo. We acknowledge Ilias Fountoulakis and Melina Zempila for performing the measurements at the Thessaloniki station and for providing the data. We acknowledge Peter den Outer for providing the Brewer data of the RIVM. Data from Uccle and Reading were obtained via Ping Wang. We acknowledge Jos van Geffen and Michiel van Weele for their stimulating discussions on various UV-related topics.

References

Greuell, W., Meirink, J.F., Wang, P. (2013) Retrieval and validation of global, direct, and diffuse irradiance derived from SEVIRI satellite observations. *J. Geophys. Res: Atm.* **118**: 2340-2361

McKinlay, A., Diffey, B.L. (1987) A reference action spectrum for ultra-violet induced erythema in human skin. In: W.F. Passchier & B.F.M. Bosnajakovic (eds.), *Human Exposure to Ultra-violet Radiation: Risks and Regulations*, Amsterdam: Elsevier, 83-87

World Health Organization (2002) Global solar UV index: A practical guide. ISBN: 92-4-159007-6

Vanicek, K., Frei, T., Litynska, Z., Schmalwieser, A. (2000) UV-Index for the public. A guide for publication and interpretation of solar UV Index forecasts for the public prepared by the Working Group 4 of the COST-713 Action "UVB Forecasting". *COST-713 Action, European Commission, Luxembourg*, ISBN: 92-828-8142-3