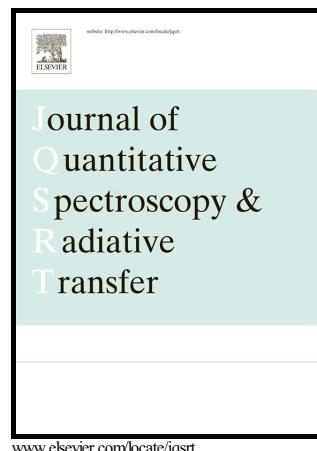


Author's Accepted Manuscript

Modeling the relationship between photosynthetically active radiation and global horizontal irradiance using singular spectrum analysis

Melina-Maria Zempila, Michael Taylor, Alkiviadis Bais, Stelios Kazadzis



PII: S0022-4073(16)30112-1

DOI: <http://dx.doi.org/10.1016/j.jqsrt.2016.06.003>

Reference: JQSRT5352

To appear in: *Journal of Quantitative Spectroscopy and Radiative Transfer*

Received date: 19 February 2016

Revised date: 15 April 2016

Accepted date: 1 June 2016

Cite this article as: Melina-Maria Zempila, Michael Taylor, Alkiviadis Bais and Stelios Kazadzis, Modeling the relationship between photosynthetically active radiation and global horizontal irradiance using singular spectrum analysis *Journal of Quantitative Spectroscopy and Radiative Transfer*, <http://dx.doi.org/10.1016/j.jqsrt.2016.06.003>

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting galley proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain

Modeling the relationship between photosynthetically active radiation and global horizontal irradiance using singular spectrum analysis

Melina-Maria Zempila^{1,2}, Michael Taylor^{2*}, Alkiviadis Bais² and Stelios Kazadzis³

¹ Natural Resource Ecology Laboratory, Colorado State University, Fort Collins, Colorado 80523, U.S.A.

² Laboratory of Atmospheric Physics, Aristotle University of Thessaloniki, PO Box 149, 54124, Thessaloniki, Greece.

³ Physikalisch-Meteorologisches Observatorium Davos, World Radiation Center, Dorfstrasse 33, 7260 Davos Dorf, Switzerland.

Correspondence to: Michael Taylor (mtaylor@auth.gr)

Abstract. We report on the construction of generic models to calculate photosynthetically active radiation (PAR) from global horizontal irradiance (GHI), and vice versa. Our study took place at stations of the Greek UV network (UVNET) and the Hellenic solar energy network (HNSE) with measurements from NILU-UV multi-filter radiometers and CM pyranometers, chosen due to their long ($\approx 1M$ record/site) high temporal resolution (≈ 1 minute) record that captures a broad range of atmospheric environments and cloudiness conditions. The uncertainty of the PAR measurements is quantified to be $\pm 6.5\%$ while the uncertainty involved in GHI measurements is up to $\approx \pm 7\%$ according to the manufacturer. We show how multi-linear regression and nonlinear neural network (NN) models, trained at a calibration site (Thessaloniki) can be made generic provided that the input-output time series are processed with multi-channel singular spectrum analysis (M-SSA). Without M-SSA, both linear and nonlinear models perform well only locally. M-SSA with 50 time-lags is found to be sufficient for identification of trend, periodic and noise components in aerosol, cloud parameters and irradiance, and to construct regularized noise models of PAR from GHI irradiances. Reconstructed PAR and GHI time series capture $\approx 95\%$ of the variance of the cross-validated target measurements and have median absolute percentage errors $< 2\%$. The intra-site median absolute error of M-SSA processed models were $\approx 8.2 \pm 1.7 \text{ W/m}^2$ for PAR and $\approx 9.2 \pm 4.2 \text{ W/m}^2$ for GHI. When applying the models trained at Thessaloniki to other stations, the average absolute mean bias between the model estimates and measured values was found to be $\approx 1.2 \text{ W/m}^2$ for PAR and $\approx 0.8 \text{ W/m}^2$ for GHI. For the models, percentage errors are

well within the uncertainty of the measurements at all sites. Generic NN models were found to perform marginally better than their linear counterparts.

Keywords: Photosynthetically Active Radiation, Global Horizontal Irradiance, Neural Networks, Singular Spectrum Analysis, Radiative Transfer, CM Kipp & Zonnen, NILU-UV

1 Introduction

Photosynthetically active radiation (PAR) occupies the narrow spectral range ($\approx 400\text{-}700\text{nm}$) of solar irradiance and is a key variable in ecophysiological models where accurate estimates are critical (Aguiar et al., 2012). PAR is used to estimate leaf photosynthesis (McCree, 1972; Alados-Arboledas 1996), to measure the productivity of forests (Landsberg & Waring, 1997), to probe the structure of forest canopies (Perry et al., 1988), and to calculate the euphotic depth of the oceans (Kirk, 1979). Moreover, the need to better constrain agents involved in climate change, mean that accurate PAR measurements have become central to reliable determination of the level of radiation intercepted by high volume biomass and its forcing effect (Zhang et al. 2000; Xu et al., 2003), as well as the impact of deforestation (Costa et al., 2007) and climate change on agriculture (Ma et al., 2012; Pei et al 2013). In terms of PAR measurement, optical instruments have been the most precise and direct method for obtaining PAR values (Udo and Aro 1999; Ross and Sulev 2000) but are not routinely carried out at the vast array of global radiometric sites (e.g. Ge et al., 2011; Wang et al., 2013; 2015a; 2015b). As a result, in many areas strongly affected by biomass burning such as the vast Amazon region of South America or the southern African savannahs for example, few empirical studies have taken place (Tsubo & Walker 2002; Aguiar et al., 2012). Consequently, PAR estimates are derived from radiative transfer models (RTMs) or from application of empirical relationships which usually depend sensitively on local conditions (Gonzalez & Calbo 2002; Gueymard, 2008; Lu et al., 2011; Aguiar et al., 2012; Wang & Shi, 2010; Janjai & Wattan, 2011; Ge et al., 2011). As a result, the advice is to use conversion relationships when combining solar measurements and meteorological parameters (Udo & Aro, 1999; Ross & Sulev, 2000; Gueymard, 1989; Bai, 2012).

The strong dependence of PAR on global horizontal irradiance (GHI) levels (Bai, 2012) means that it can be derived from a potentially large pool of global measurements. Furthermore, due to the outstanding growth of technologies targeted at

deploying the potential of solar energy for electricity production (IEA, 2012; IEA, 2014: Executive Summary) as an abundant renewable resource (Szurovi et al., 2007); Burnett et al. (2014); Sarralde et al. (2015)), a large number of studies (Fernández-Peruchena & Gastón, 2016; McCandless, Haupt & Young, 2015; Zell, 2015; Sundaram & Sarat Chandra Babu, 2015; Pierro et al., 2015; Pedro & Coimbra, 2015; Zempila et al., 2016a; Kashyap, Bansal & Sao, 2015; Lopez et al., 2001) have proposed many methods to derive the GHI that is exploited in Photovoltaic systems (PV) (Salas et al, 2009), and to assess, map, validate and predict its abundance worldwide.

While PAR and GHI are strongly inter-connected, a lack of coincidence in their measurements results in inhomogeneous time series that may contain large data gaps. In addition, the PAR to GHI ratio experiences large diurnal or/and seasonal variations due to cloud fraction, solar elevation and perceptible water in the atmosphere (Alados et al, 1996; Udo & ARO, 1999; Ross & Sulev, 2000). Thus, different models that assess PAR and GHI under the whole range of sky conditions can be found in the literature (Hanan & Begue 1995; Alados et al., 1996; Ross et al., 1998; Gueymard, 1989; Alados-Arboledas et al., 2000; Ross & Sulev, 2000; Zhang et al., 2000; Cho et al. 2003; Rubio et al., 2005; Bosch et al. 2008, Bai, 2012).

In this work, we therefore develop and assess the performance of linear regression (LR), multiple linear regression (MLR) and nonlinear neural networks (NN) to generically calculate PAR from GHI (and vice versa). In the multivariate models we include the solar zenith angle (SZA), the columnar perceptible water vapour (WV) and the aerosol optical depth (AOD) at 500nm in order to account implicitly for the combined effect of cloud and aerosol.

The rest of this report is arranged as follows. In Section 2 we describe the study region and overview the general characteristics of the substantive record of calibrated measurements from UVNET and HNSE used in this work. In Section 3, we carefully describe the data pre-processing and instrumental techniques applied to ensure the integrity of the time series used for model development and assessment. Multi-channel singular spectrum analysis (M-SSA) which we use for decomposition, noise regularization and reconstruction of the time series is described here. Section 4 then presents the construction of linear and nonlinear NN models of PAR from GHI (and vice-versa) together with a description of the procedure used to train and optimize them. Here, we also report their range of validity. In Section 5, we present simulation results for PAR and GHI obtained from feeding the models with satellite-calibrated inputs, and we discuss to what extent the models are generic. We sum up our findings in Section 6 and suggest further work and potential areas of application.

2 Study area and observations

We have selected Greece as the study region for two principal reasons: i) a substantive record (>100K measurements per monitoring site) of high temporal resolution (\approx 1 minute) co-located data is provided by the two monitoring networks (PAR from UVNET and GHI from HNSE), ii) the region (latitude: 34°-42°N, longitude: 19°-28°E) is located at a crossroad of seasonal aerosol flows (influenced for example by dust episodes from Northern Africa and the Middle East as well as local and transcontinental clouds of biomass burning products) and is exposed to a broad range of atmospheric and cloud conditions. Greece (latitude: 34°-42°N, longitude: 19°-28°E) is located at the southeastern end of the European continent and is characterized by a mountainous peninsular mainland in the southeast Mediterranean basin, covering a region of nearly 132,000 km².

2.1 UVNET: NILU-UV PAR measurements

In line with the scientific objectives of UVNET, 7 monitoring stations were established during 2005 to monitor solar irradiance under different atmospheric environments (urban and rural) and climatological conditions. UVNET has the necessary infrastructure in place to ensure smooth and seamless operation and network interconnectivity of the UV monitoring instruments and, in particular, each station is equipped with a NILU-UV multi-filter radiometer that provides irradiance measurements in 5 UV wavelength bins (nominally at 305, 312, 320, 340 and 380nm) together with a sixth channel for measuring PAR. For all channels, mean values are recorded at one-minute intervals together with the corresponding standard deviation.

2.1.1 Absolute calibration

The absolute calibration of the PAR channel of the NILU-UV radiometers is easy to achieve by comparison against a calibrated standard lamp. While this is available at LAP, it is not available at the other stations. Therefore an alternative, slightly less accurate method has been used based on simulations with the libRadtran radiative transfer model (Mayer & Kylling, 2005) and following the methodology proposed by Dalhback (1996). In particular, irradiance spectra from 290-4500 nm were derived for each station of the network with the *uvspec* model based on the following input information: the total ozone column (TOC), the AOD at 550nm, WV, SZA, the date, and the height above sea level. For all stations used in this

study, TOC, AOD and WV were derived from satellite overpasses. In particular, daily overpass data (Collection 3, V8.3) from the ozone monitoring instrument (OMI) onboard the Aura platform was used for TOC and daily data gridded at $1^\circ \times 1^\circ$ (Collection 5, Level 3) from the moderate resolution imaging spectroradiometer (MODIS) onboard the Aqua and Terra platforms (also in the A-train), was used to retrieve AOD and WV. Spectra resulting from *uvspec* simulations were then evaluated against concurrent clear-sky measurements of the GHI derived from the CM type Kipp & Zonnen pyranometers of the HNSE (see Section 2.2 for details). After the evaluation and spectral refinement of the RTM dataset based on comparisons made with ground-based GHI measurements (differences of 2% and below were obtained during the performance of this quantitative test), PAR measurements were then calibrated by taking into account the trend (corrected by smoothing spline implementation and found to be of the order of $< 3\%$) and the seasonality corrected by 3-cycle Fourier analysis using the Levenberg–Marquardt algorithm (LMA) and a least absolute residual (LAR) robust fit to minimize the effect of sparse extreme values due to random incidences that were not monitored by the quality control imposed on the data series. The seasonal component was detected to be $< 5\%$. Thus, any small degradation of the PAR channel and the effects of azimuthal and cosine errors were implicitly taken into account, while the error due to the non-ideal spectral response of the PAR channel was also implicitly corrected through model assimilations that were performed with an ideal flat weighting function.

Furthermore, in Thessaloniki, where a double-monochromator Brewer spectroradiometer and a Cimel sunphotometer, as part of NASA’s aerosol robotic network (AERONET) are in regular operation, TOC, AOD and WV from these ground-based instruments were also used as input data in the RTM simulations. Mean differences due to the different types of input data used to calibrate the PAR data, were found to be less than 0.5% with a standard deviation of 1.21%. Consequently the calibration of the PAR channel using satellite inputs in the model-derived estimates can be considered to be of comparable accuracy to that obtained with ground-based input data. For all 6 NILU-UV multi-filter radiometers, the absolute percentage difference between the PAR data and RTM estimates is provided in Table 1 together with the associated number of observations. In Thessaloniki, both the ground-based (GRD) and satellite-based (SAT) calibration procedures are evaluated to assure the homogeneity of the method applied to all 6 stations.

Table 1 Absolute calibration percentage error and the corresponding number of concurrent measurement observations from the PAR channel of the 6 participating NILU-UV stations used in this study. The absolute error for the co-located satellite calibration at Thessaloniki is also shown.

UVNET Stations	Abbrev.	Absolute Error (%)	No of Observations (N)
Athens	ATH	± 1.26	298
Finokalia	FIN	± 0.67	77
Ioannina	IOA	± 1.88	320
Mytilene	MYT	± 1.15	480
Thessaloniki (GRD)	THE	± 2.04	3418
Thessaloniki (SAT)	THE	± 1.66	428
Xanthi	XAN	± 1.41	199

In order to investigate the level of accuracy of the absolute calibration factor derived in this way, data calibrated with the original calibration certificate of NILU-UV 04103 were compared and differences between the two calibrations were calculated for a day that was cloud-free (6th of June 2005, 17.2<SZA<-80) and a day when clouds were present (8th of May 2005) throughout the whole sunshine duration. Under clear skies (cloud free cases) the percentage differences were found to be equal to ~3.73%, under scattered cloud conditions the estimated bias was 5.26%, while under overcast skies with low irradiances the error reached a maximum of 8% when the RTM-derived calibration was found to overestimate the derived PAR values. Since the manufacturer provides the PAR data in $\mu\text{E}/\text{m}^2/\text{s}$ an additional 3.97% of uncertainty is also calculated based on each RT model spectrum used based on the conversion number provided by Young et al., 1987. Therefore the overall uncertainty of the absolute calibration of the PAR channel is equal to 11.72-14.3% or $\pm 6.51\%$ (section 2.1.4).

2.1.2 Optical characterization

Prior to deployment in UVNET stations, all instruments were optically characterized at the facilities of the Laboratory of Atmospheric Physics (LAP) in Thessaloniki, Greece, and the spectral, angular and azimuthal responses are calculated for each channel. The optical characterization unit consists of a 1000W Xenon lamp (ORIEL model 6269) with output spectral range between 220-2400nm that serves as the radiation source. A double monochromator (Jobin-Yvon SpectraMate 1680) with a triangular slit function of FWHM = 2.25 nm produces quasi-monochromatic radiation. The monochromatic beam is separated into two components by a quartz beam splitter that transmits 90% to the tested instrument and reflects 10% to a calibrated photodiode (Gigahertz-Optik PD-9306) that monitors the stability of the monochromator's output with an uncertainty of $\pm 4\%$ in the spectral region 380-900nm. By combining the individual uncertainty levels from the key components of the optical characterization unit, we estimate that the uncertainty in determining the spectral response of the PAR channel is ~5%. To be specific, with uncertainties associated with the photodiode, the monochromator's wavelength

accuracy and the stability of the monochromator's output being $\pm 2\%$, $\pm 0.16\%$ and $\pm 4\%$ respectively, the combined uncertainty is calculated to be $\sqrt{[(2\%)^2 + (0.16\%)^2 + (4\%)^2]} = \pm 4.475\%$.

For the angular response, determination, the same unit is used while the tested instrument is rotated around a vertical axis perpendicular to the output beam of the monochromator. To increase the signal at large incidence angles, the gratings in the monochromator are set to non-dispersive position so that the whole spectrum of the lamp is directed to the output. The uncertainty of this measurement is only affected by the accuracy of the rotation system and the stability of the photodiode, and is estimated to be of the order of $\sim 5\%$. To be specific, here we summed the accuracies associated with the photodiode (2%), the positioning of the instrument (2.5%) and the rotation of the system (0.01 degrees over each 1 degree angular interval = 1%) giving a total of 5.5% . In Fig. 1 the spectral response of the PAR channel of each of the 6 UVNET NILU-UV instruments used in this study are provided. The spectral response increases almost monotonically with increasing wavelength with the value at 400 nm being about 20% of the maximum that occurs at around 700 nm and is far from the expected ideal response of unity over this wavelength range. This deviation from the ideal response results in a small dependency on SZA and AOD variations of up to about 1.5% and 2% , respectively, for all SZAs, and for AOD at 500 nm between 0 and 1.2 . Changes of up to $\pm 0.5\%$ in the integral of the spectral response of the NILU-UV 04103 PAR channel were detected from two measurements taken 5 years apart, while the general shape of the response remained practically the same.

PAR Spectral Response

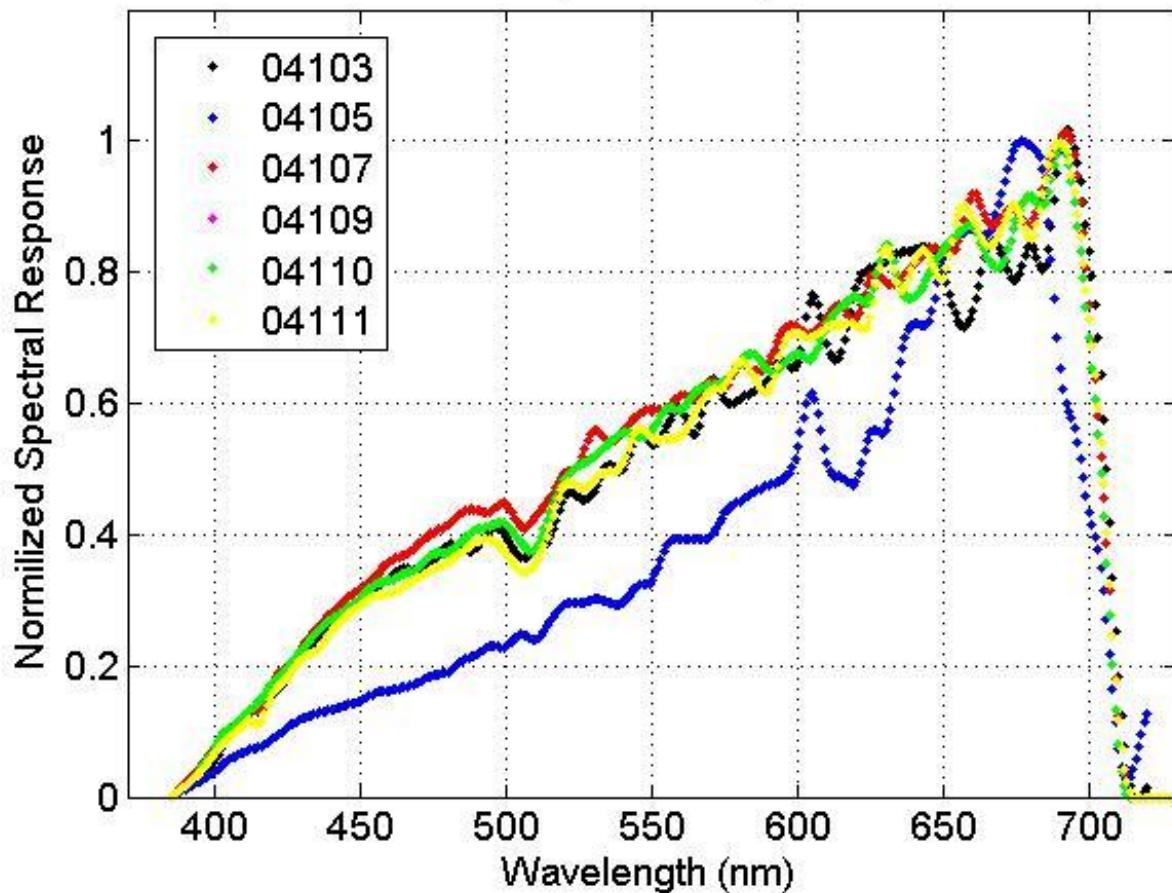


Figure 1 PAR channel spectral responses (normalized to the maximum) for the 6 NILU-UV instruments of UVNET.

In Fig. 2 the cosine error of the PAR channel is shown for NILU 04103 (the error for the other instruments are similar and are not shown for the sake of clarity). The cosine error of the PAR channel is found to be < 10% for angles of incidence < 80° for all 6 instruments tested.

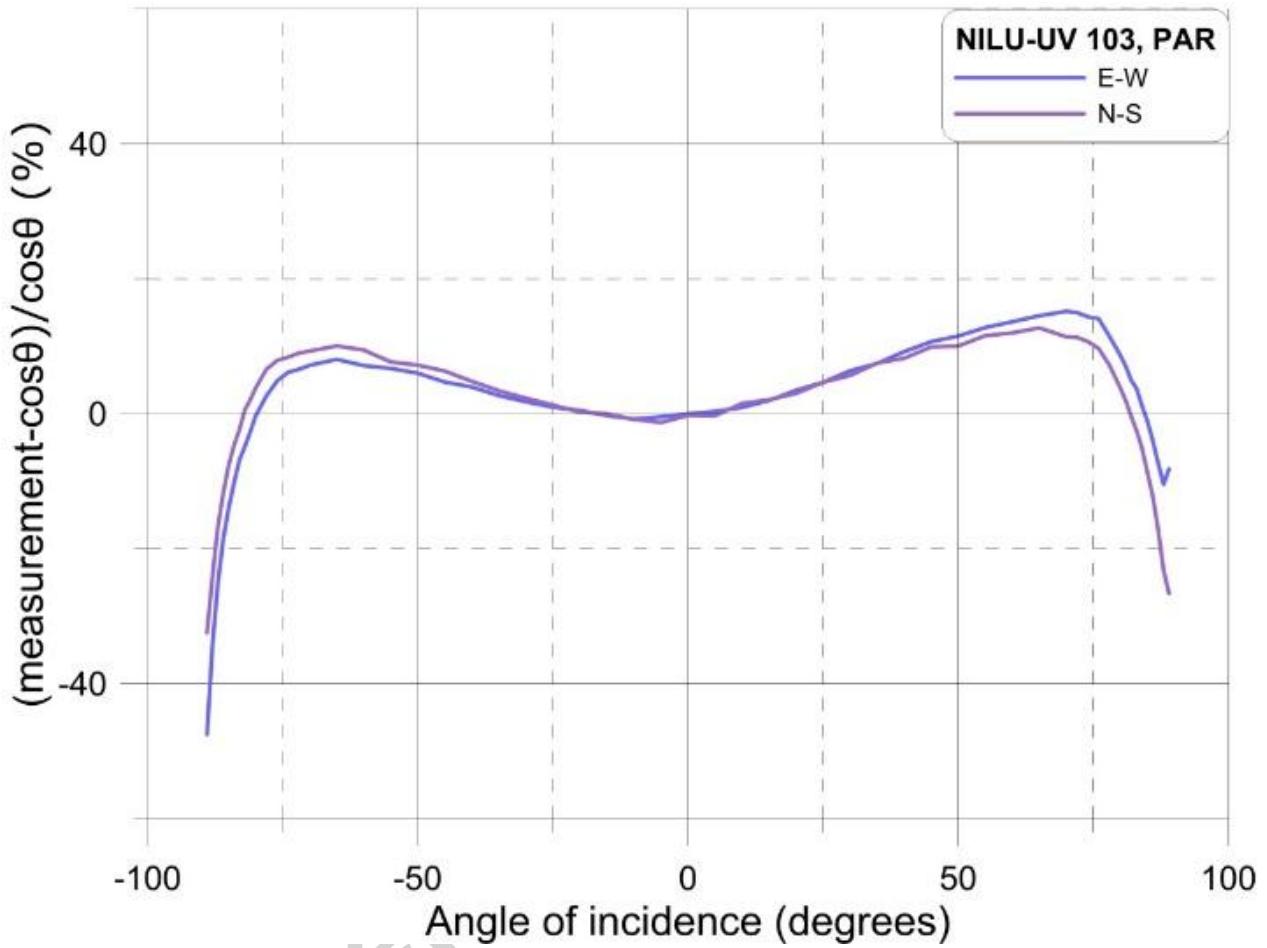


Figure 2 Cosine error of the PAR channel of NILU-UV 04103 in two perpendicular azimuthal directions.

Although angular response corrections are not typically applied on NILU-UV PAR measurements (Lakkala et al, 2016), in our case the cosine error of the PAR channel was estimated from the measured angular-response data in accordance with standard practices (Seckmeyer et al., 1993; Feister et al., 1997, Landelius et al., 2000). In order to derive the contribution of the direct and diffuse radiation to the cosine correction factor, knowledge of the ratio of direct to global irradiance is necessary (Bais et al, 1998). For PAR, the contribution of the direct component (which is directly proportional to the cosine error), is comparable to that of the diffuse component due to the reduced scattered radiation in this wavelength band, and was found to range on average between about 60% and 25% for SZA of 20° and 70°. Taking into account that the cosine correction for the direct component is very small for small SZAs (up to 3% for SZA = 20°) and increases to 15% at SZA = 70° (see Fig. 2), the contribution of the direct component to the correction factor is of the order 2-4%. The contribution of

the diffuse component can be quantified through the f_d parameter given in Eq. 1 under the assumption that diffuse radiation is isotropic (Bais et al, 1998):

$$f_d = \frac{\iint R_j(\theta) \cdot \sin\theta \cdot \cos\theta d\theta d\varphi}{\iint \sin\theta \cdot \cos\theta d\theta d\varphi} \quad (1)$$

where: θ is the solar zenith angle, φ is the azimuthal angle, and $R_j(\theta)$ is the cosine response of the j^{th} channel. The f_d derived for NILU-UV 04103 from the mean of the angular spectral response measured at two perpendicular azimuthal directions was estimated to be $\pm 3.5\%$, resulting in a contribution of the diffuse component to the cosine correction factor of 2-4% for SZAs between 20° and 70° . Therefore, the cosine correction factor for PAR under clear skies is 4-8% and is about 4% under cloudy conditions. Similar cosine errors are calculated for the other six instruments of UVNET. Finally, in this sub-section, the azimuthal dependence of the PAR channel was further evaluated by sequential measurements under real conditions for different azimuthal orientations of the instrument at an approximately constant SZA of 60° . The measurements at different azimuths were found to have a variability of 2.7% and are well below the limit of 5% set by other studies (e.g. Høiskar et al., 2003).

2.1.3 Stability Tests

Although the instruments are temperature controlled to $50^\circ \pm 0.005^\circ \text{C}$, the temperature effect on PAR measurements from NILU-UV 04103 was investigated based on lamp measurements. The raw counts of the PAR channel were monitored for different internal temperatures in the range 35° - 55° C and showed insignificant sensitivity ($0.002\%/\text{ }^\circ\text{C}$) to temperature changes. The linearity of the channel is also investigated during the optical characterization in the indoor unit by changing the monochromator's white light output with lenses. No linearity issues were detected for the PAR channel. Moreover, the instruments were checked daily for the level of their dark signal making sure that this signal was at least 10^4 times lower than the magnitude of the signal during daytime. The long-term stability of the PAR channel is monitored routinely on a monthly basis in Thessaloniki, using alternatively two sets of three 150W lamps. The current of each lamp is monitored by a computerized system and is stabilized to 6.6 A. Each lamp is warmed for about 5 minutes before it is placed onto the instrument to avoid overheating the diffuser. Since this increases the risk of damaging the lamp, extra care is taken to smoothly move and position the lamp. The measurement lasts for 1 minute and the lamp is then removed. After 1 minute the

same lamp is measured once again aiming to exclude any random and uncontrolled incidences during the lamp measurements. For monitoring the stability of the PAR channel, mean values of the measured raw data are used. All 6 lamps result in a quite stable signal and their maximum percentage coefficient of variation (standard deviation relative to the mean over all 6 lamp values) during the under study period is ~3.1%. Furthermore for NILU-UV 04103 in Thessaloniki, comparison against a very stable CM-21 pyranometer of Kipp & Zonen (Bais et al, 2013) (see section 2.2) did not reveal any long-term trend in the ratio PAR counts over GHI calibrated data, apart from a seasonal component attributable mainly to the seasonal cycle of water vapor that affects almost exclusively the GHI measurements. Similar comparisons were performed for all stations included in this study so as to assess the stability of the PAR channel for the period investigated at each station.

2.1.4 Overall uncertainty of the PAR data

In Table 2 the uncertainties of each key component of the PAR calibration is evaluated and the overall uncertainty of this dataset is assessed. We have assumed that errors related to tilt are negligible for this type of silicon diffuser. Furthermore, the NILU-UV instrument in Thessaloniki has been routinely checked during the ten year period of its operation and it has not been necessary to address any changes in the tilting even though severe weather conditions (strong winds like the Vardaris flow) are dominant at the greater area of Thessaloniki. Here, we also wish to note that the study period excludes the first 5 months of deployment in order to minimize the effect of severe degradation (particularly the impact of soiling on the UV channels) associated with the initial exposure of the instrument to real weather conditions. In UVNET therefore, our instruments are maintained with caution especially to the Southern part of Greece were Sahara dust events are more frequent (e.g. Finokalia, Crete).

Table 2 Uncertainties of each key component of the calibration procedures applied to the PAR data.

Type	Estimated Uncertainty
Optical Calibration unit	
Spectral response	5%
Angular response	5%
Change of spectral response/year	0.2%
Calibration Procedure	
Absolute calibration	6.91%
Angular response	4-8%
Spectral response	5%
Azimuthal response	2.7%
Variability in the model estimations	2.88%
Overall Calibration Procedure	
	11.72-14.3%

The application of the RTM is a component that introduces an additional uncertainty of 5.99% (absolute calibration and variability in the RTM estimations). Its necessity is due to lack of independent absolute calibration at stations that are not monitored routinely via lamp checks or other ancillary data. Based on Table 2, we are able to assure that the accuracy of the time series obtained from the PAR channel is on average $\pm 6.5\%$. The validity of this calibration procedure is further verified (indirectly) by successful deployment of the models on data from other instruments in UVNET that have unique correction and calibration factors (see section 5.3). We summarize the steps involved in calibration of the PAR channel in Fig. 3.

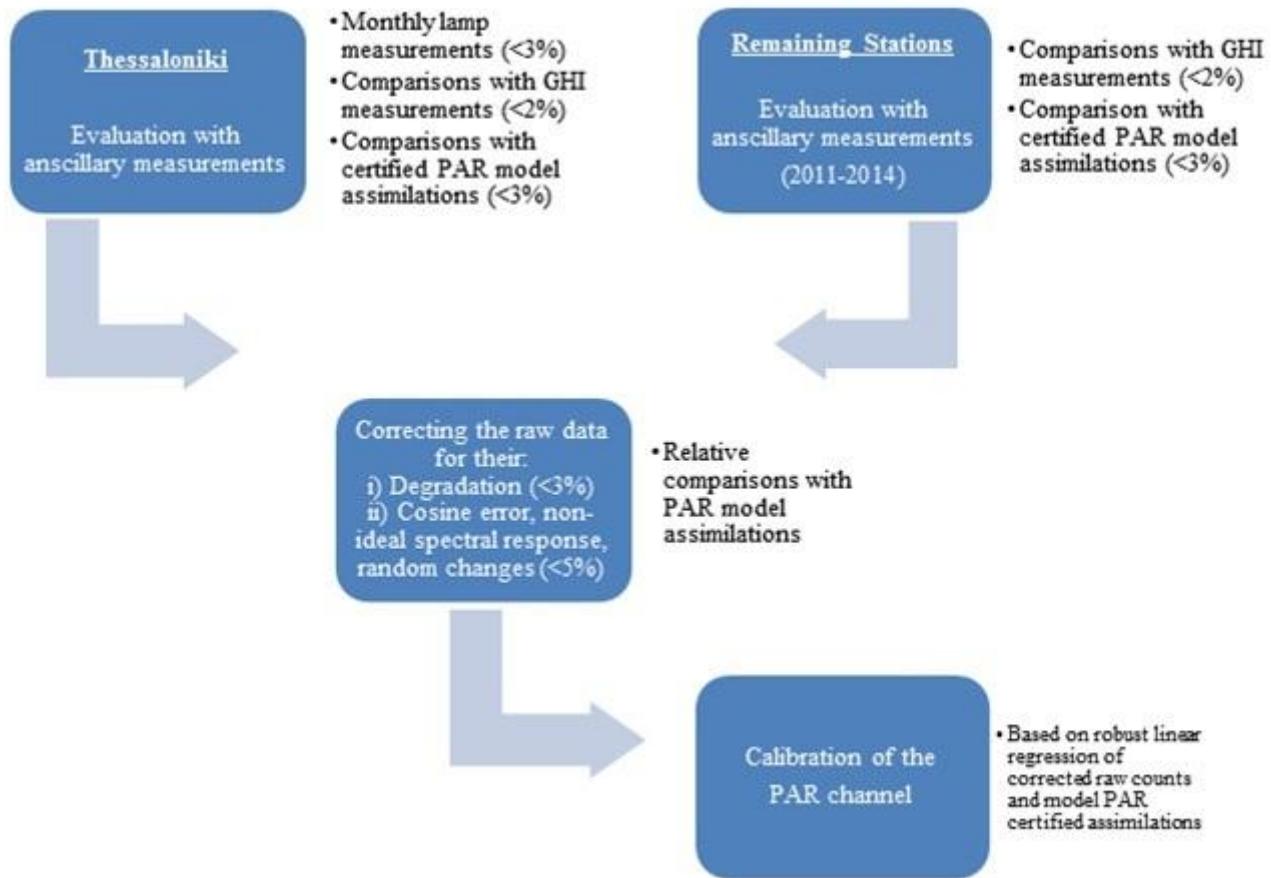


Figure 3 Schematic of the key procedures used to calibrate the PAR channel of NILU-UV at all 6 stations of UVNET.

2.2 HNSE: CM measurements

In 2011, the HNSE was established to augment UVNET (Bais et al., 2013), and extended the number of monitoring stations from 6 to 13 to increase the spatial representivity of measurements in the region. Interestingly, a cluster analysis of the mean cloud coverage over Greece performed by Zagouras et al. (2013, 2014) suggests that the country can be divided into 22 spatially-distinct regions, each with their own characteristic (mean) cloud conditions. This is relevant to an assessment of the spatial representivity of the monitoring network. Each station of the HNSE is equipped with CM (Kipp & Zonen) pyranometers (types CM10, CM11 and CM21), that provide GHI records at one-minute intervals as mean values together with the corresponding standard deviation (usually between 50 and 60 seconds). Geolocation and Pyranometer type information is summarized in Table 3.

Table 3 Geolocation and pyranometer type information for instruments used to provide GHI measurements at the monitoring stations of HNSE used in this study. Supplementary information for the NILU-UV instruments of UVNET is also provided for completeness.

Station	Abbrev.	Latitude (°N)	Longitude (°E)	Altitude (m)	Pyranometer	NILU-UV model #
Athens	ATH	37.99	23.78	180	CM21	04105
Finokalia	FIN	35.34	25.67	250	CM11	04109
Ioannina	IOA	39.62	20.85	541	CM10	04111
Mytilene	MYT	39.11	26.55	80	CM11	04107
Thessaloniki	THE	40.69	22.96	60	CM21	04103
Xanthi	XAN	41.14	24.89	75	CM11	04110

Prior to deployment in HNSE stations, all pyranometers were calibrated in the laboratory via the standard technique of comparing the GHI to that obtained from a 150 Watt Hg radiation source. Absolute calibration factors were deliberated and applied to the raw dark-corrected signals in order to assess the GHI at each station. Recalibration of the instruments is scheduled for repetition every 5 years in accordance with the documented stability of CM pyranometers (Bais et al., 2012). According to the manufacturer's manual, CM typepyranometers may suffer from nonlinearity ($<\pm 0.2\%$), spectral selectivity ($\pm 2\%$), temperature dependence of sensitivity ($<\pm 1\%$), tilt errors ($<\pm 0.2\%$), instability ($<\pm 0.5\%$) and a cosine error of the order of $\pm 6\%$. Thus we can conclude that in general the maximum uncertainty for the GHI measurements is $\sqrt{[(0.2\%)^2 + (2\%)^2 + (1\%)^2 + (0.2\%)^2 + (0.5\%)^2 + (6\%)^2]} = \pm 6.43\%$.

As has been pointed out by Gueymard and Myers (2008) in the context of the construction of new models, data is often measured at one or a small number of specifically chosen sites, and the generality of a given model is then verified by tests performed at other sites. This is the approach we adopt here. In particular, since measurements at the Thessaloniki node are well-characterized, have documented uncertainties, are quality-assured, volumous and of high temporal resolution, this site will be used for training and validating the GHI and PAR models developed in this study - so that they can be tested with confidence elsewhere. Also, with careful emphasis on the min-max ranges of parameters used during the model development stage, we will aim to capture as broad a range of atmospheric conditions as possible and hence to maximize the generality of the models.

3 Preliminary data analysis and pre-processing

3.1 Data redundancy and co-linearity tests

In Section 4 we will describe the construction of LR, MLR and NN models that relate a target (output) variable (PAR or GHI) to a set of input variables (SZA, WV and/or AOD) plus either PAR or GHI depending on the choice of target variable. While a physical basis for this choice of variables has been described in Section 1, we began by assessing whether or not there was any possible redundancy in the choice of 4 inputs variables using principal components analysis and factor analysis (PCA-FA). The aim of this is to see if it is possible (and/or necessary) to reduce the complexity of the model and any potential co-linearity, and determine the relevant variables needed to derive the values of the output variable from a candidate list of input variables. The PAR and GHI models we construct have the following functional forms:

$$\text{PAR} = f_1(\text{GHI}, \text{SZA}, \text{WV}, \text{AOD}) \quad (2)$$

$$\text{GHI} = f_2(\text{PAR}, \text{SZA}, \text{WV}, \text{AOD}) \quad (3)$$

and the sought-after candidate list of input variables is provided by the right-hand sides of the above two equations. We performed PCA-FA and found, for the input variable matrices for the PAR and GHI models of Eq. 2 and Eq. 3, the correlations between the input variables are not strong enough to warrant or require dimensionality reduction of the data, and that working with PCs instead of the original variables does not appear to provide any statistical advantage. The corollary of this is that PCA shows that the four inputs (SZA, WV, AOD and irradiance) are strongly linearly independent and therefore,

given the physical basis outlined in Section 2, are suitable for model development. The details of the full PCA-FA can be found in Appendix A.

3.2 Empirical conversion of PAR counts to irradiance and satellite calibration

In preparing time series for this study, we converted raw PAR counts to irradiance estimates in units of W/m^2 so as to facilitate a direct comparison with GHI irradiances. The empirical approach we adopt, described in detail below, is motivated by a recent study demonstrating that light intensity loggers for estimating PAR have been found to be as accurate and reliable to within 3.8% as factory-calibrated scalar PAR sensors under a wide range of environmental conditions (Long et al., 2012).

In order to indirectly calibrate the PAR measurements, clear sky raw counts (using the cloud detection algorithm proposed by Zempila et al., 2016b) were first regressed on RTM estimates obtained with satellite data as inputs, but a clear linear bias was detected that increases with the size of the PAR irradiance as shown in Fig. 1. Although the magnitude of the trend and seasonal component due to degradation and cosine and azimuthal errors are less than 3% and 5% respectively, detrending via application of a smoothing spline interpolation in conjunction with deseasonalization of the raw PAR counts using 3-cycle Fourier analysis was able to remove this bias (see Fig. 4) and produce a clean linear mapping between counts and RTM model estimates having an intercept of $\approx 3.7 \text{ W/m}^2$ and a slope of 1.009 (Pearson R=0.987).

Thessaloniki, Satellite Derived Calibration

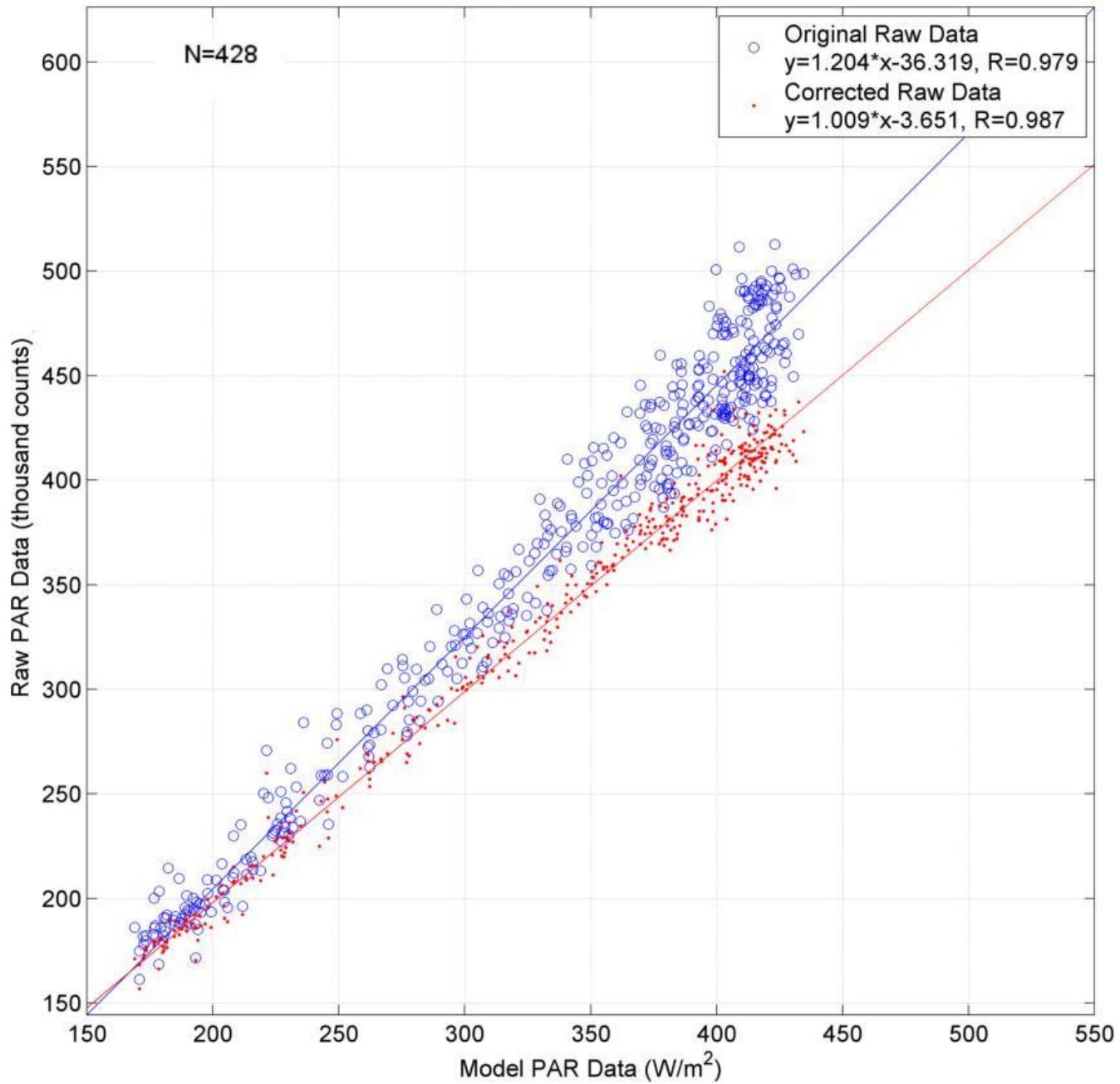


Figure 4 The effect of detrending and deseasonalizing on the linear regression of clear skies raw PAR count data on RTM estimates obtained using satellite data as inputs. The plot of raw data is shown by blue circles while the plot of detrended and deseasonalized data is shown by red dots. All data is from the Thessaloniki site in the HNSE.

3.3 Time series regularization with multi-channel singular spectrum analysis (M-SSA)

The time series data used in this study, being of geophysical origin, are highly variable. The SZA, PAR and GHI signals are dominated by the diurnal cycle. Furthermore, the irradiances are strongly modulated by sporadic and non-deterministic changes in cloud cover, columnar WV and AOD which also have an impact on surface radiation levels (Taylor et al., 2016). As such, the time series are very noisy (we will address the issue of whether or not they are linearly or nonlinearly deterministic in Section 5). In order to facilitate NN learning of the functional relation f_1 and f_2 in the PAR and GHI models of Eq. 2 and Eq. 3, the time series variables were made covariant (i.e. normalized by centering on the mean and scaling to the standard deviation), and were detrended and deseasonalized using singular spectrum analysis (SSA). Before presenting the details of SSA, we would like to emphasize that this type of detrending and deseasonalization is independent of the methodology applied during the calibration procedures for the PAR measurements, and is not connected in any way to the accuracy of the time series themselves. The rationale behind performing this type of time series decomposition is to be able to correctly model the effects of WV and AOD which can be masked in the daily pattern of the irradiance data, especially during clear sky conditions and on days with light broken/scattered cloud. On overcast days, or obscuration events associated with rapid variations in cloud density, the daily curve is probably nearly 100% noise. On these particular days, SSA does not provide that much of advantage.

Being based on the residual time series, the NN models therefore seek the underlying relation between the covariant *noise* signal for the output irradiance and that of each input variable. The modeled PAR and GHI signals are then reconstructed by re-incorporating the extracted trend and periodic components and denormalizing the reconstructed output signal. While it may be surprising to want to perform function approximation on the residuals, one should note that noise (usually Gaussian) is often deliberately added to NN models to help regularize highly variable input data and prevent them from yielding unreasonable responses (Holmström & Koistinen, 1992; Bishop, 1995; Di Noia et al., 2015). Here, we are doing the exact opposite, i.e. we are removing the deregularizing trend and periodic components from the input and output time series, and then modeling the noise. The results of Section 5 suggest that this approach is effective for the LR, MLR and NN models we construct. We shall also see why M-SSA is key to the generic capabilities of the (nonlinear) NN models and their (linear) LR and MLR counter-parts. For now, we outline the practical details of the M-SSA approach we have adopted to decompose the

input and output time series into additive trend, periodic and noise components, and to produce the regularized noise models used in this study.

SSA is a non-parametric spectral estimation technique that can separate time series into a sum of interpretable components (see the two reviews by Golyandina et al., 2001; Ghil et al., 2002 for more mathematical detail). The method gets its name from the spectrum of eigenvalues in a singular value decomposition (SVD) of a lag-covariance matrix formed from M successive lagged versions of the time series. The lag-covariance matrix can be estimated directly from the data as a Toeplitz matrix with constant diagonals (Vautard & Ghil, 1989) and can lead to better noise reduction, but it has been noted that this can lead to a slightly larger bias when the time series is strongly non-stationary (Ghil et al., 2002). More importantly however, for long time series (e.g. $N > 10^5$ as in the case of the simulation runs we perform in Section 5), the calculation of the SVD in this way is prohibitive in terms of computer runtime on even a relatively powerful desktop PC (e.g. a quad-core 3GHz 16Gb RAM workstation running MATLAB under Windows 7). In a different approach to the problem of estimating the SVD, Broomhead and King (1986) showed how rather accurate decompositions (and therefore accurate reconstructions) of time series can be achieved by embedding the time series in a vector space of lower dimension using the Theorem of Takens (1981). In this delay-coordinate phase space, SSA untangles trajectories by providing data-adaptive filters that help separate the time series into statistically independent eigenmodes (“components”) at zero lag associated with trends, oscillatory patterns, and noise. It is important to note that trends need not be linear and that oscillations can be amplitude and phase modulated (Ghil et al., 2002). So, while nonlinear oscillations often require the use of many harmonics or subharmonics of the fundamental period when carrying out classical Fourier analysis, a handful of SSA eigenmodes might suffice to model the same dynamics (Ghil et al., 2002). For the set of $N = 47088$ coincident input-output data records at Thessaloniki we used in this work to train our models, we performed M-SSA (see Section 4.2 of Ghil et al., 2002 and also Hassani & Mahmoudvand, 2011 for details of the multivariate extension of SSA to multiple time series) on $L = 5$ time series (4 input variables plus 1 output irradiance), embedding them concurrently in a lower dimensional space with $M = 50$ delays each (this was found to be sufficient since SSA reconstruction errors were less than 2% relative to the measured data as we demonstrate below). To apply M-SSA, we first of all constructed for each coincident time series of SZA, WV, AOD, PAR

and GHI, the $N \times M$ “trajectory matrix” \mathbf{Y} formed from time-delayed versions of each time series such that, in the case of PAR time series for example,

$$\mathbf{Y}_{\text{PAR}} = \begin{bmatrix} \text{PAR}_1 & \text{PAR}_2 & \text{PAR}_3 & \cdots & \text{PAR}_M \\ \text{PAR}_2 & \text{PAR}_3 & \text{PAR}_4 & \cdots & \text{PAR}_{M+1} \\ \text{PAR}_3 & \text{PAR}_4 & \text{PAR}_5 & \cdots & \text{PAR}_{M+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \text{PAR}_{N-1} & \text{PAR}_N & 0 & \cdots & \text{PAR}_{N+M-2} \\ \text{PAR}_N & 0 & 0 & \cdots & \text{PAR}_{N+M-1} \end{bmatrix} \quad (4)$$

Note that all elements where the row index $j > N$ are equal to 0 result from padding the lag series so that their columns are of length N . Note also that the trajectory matrices are Hankel matrices since the values on the anti-diagonals are equal. The multi-channel trajectory matrix \mathbf{Y} used to apply SSA to the variables in Eq. 2 and Eq. 3, is therefore a block matrix formed by concatenating (or “stacking”) the trajectory matrix of each variable having the same form as Eq. 4 so that,

$$\mathbf{Y} = [\mathbf{Y}_{\text{SZA}} \mathbf{Y}_{\text{WV}} \mathbf{Y}_{\text{AOD}} \mathbf{Y}_{\text{PAR}} \mathbf{Y}_{\text{GHI}}] \quad (5)$$

The next step is to perform SVD on \mathbf{Y} . This is achieved by calculating the $LM \times LM$ lag-covariance matrix \mathbf{C} ,

$$\mathbf{C} = \frac{1}{N} \mathbf{Y}^T \mathbf{Y} \quad (6)$$

which has LM eigenvalues (λ_k) and eigenvectors (E_k). We sort λ in decreasing order together with the associated eigenvectors such that each eigenvalue λ_k accounts for the partial variance in the direction given by E_k and that the sum of the eigenvalues (i.e. the trace of C) gives the total variance of the original time series. The quantities $\sqrt{\lambda}$ are the singular values of C from which SSA gets its name. Note that for M-SSA, each E_k is composed of L consecutive segments of length M . Next, we produced the principal components (PC) matrix A of size $N \times LM$ by projecting the multi-channel trajectory matrix \mathbf{Y} onto the matrix of eigenvectors E :

$$\mathbf{A} = \mathbf{YE} \quad (7)$$

Since the eigenvectors are sorted, the effect of this projection is to gather most of the variance in the first PC, most of the remaining variance in the second PC and so on with the variance of each PC being identically equal to the eigenvalue of the corresponding eigenvector. In M-SSA, each element of the PC matrix A is the sum of a linear combination of M values from each time series weighted by the corresponding part of the eigenvector E . This means that each PC contains characteristics of

all time series (e.g. they are able to capture periodicities common to all variables). To reconstruct time series in the original coordinates, it is then necessary to “undo” the time-delay embedding by forming an inverse-trajectory matrix Z of size N x M for each of the PCs associated with each time series variable (SZA, WV, AOD, PAR and GHI). To do this, for each of the m=1:M components, a Z matrix is formed from that component’s PC with the time delay running in the opposite direction (i.e. a time-lead rather than time-lag):

$$Z^m = \begin{bmatrix} PC_1^m & 0 & 0 & \dots & 0 \\ PC_2^m & PC_1^m & 0 & \dots & 0 \\ PC_3^m & PC_2^m & PC_1^m & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ PC_N^m & PC_{N-1}^m & PC_{N-2}^m & \dots & PC_{N-M+1}^m \end{bmatrix} \quad (8)$$

Note that, this time, padding with zeros is at the front-end of the time-lead columns of Z. The time series reconstruction of each component of each time series variable in the original time series coordinates is then obtained by convolving its Z matrix with the eigenvector for that variable’s component:

$$RC_l^m = \frac{1}{M} Z^m E_l^m \quad (9)$$

where l = 1:L are the variables and m = 1:M are the components of each channel (i.e. input-output variable). Inclusion of all components faithfully recreates the original time series. However, we are interested in being able to interpret components as trend, periodicity and noise so as to produce stationary and regularized noise models. The key to this process are the eigenvectors that explain a large portion of the variance as measured by their eigenvalues. A scree plot of the eigenvalues of the lag-covariance matrix C versus their rank is used to identify a break point between the high variance periodic components and a shallow plateau (a “long tail”) of components that are associated with a low frequency trend and the remaining noise components. Fig. 5 shows the rank-eigenvalue spectrum for the trajectory matrix Y of Eq. 5 formed from the input-output variables used in this study. The separation between periodic components and trend and noise components is identified from the break-point where the absolute derivative of the eigenvalue curve (normalized to the maximum eigenvalue) is less than 1. The total periodicity is then obtained from the sum of the periodic components preceding the break point. The trend component is identified as that having the maximum absolute mean value of the remaining non-periodic components, and the noise signal is obtained from the sum of the non-periodic components excluding the trend.

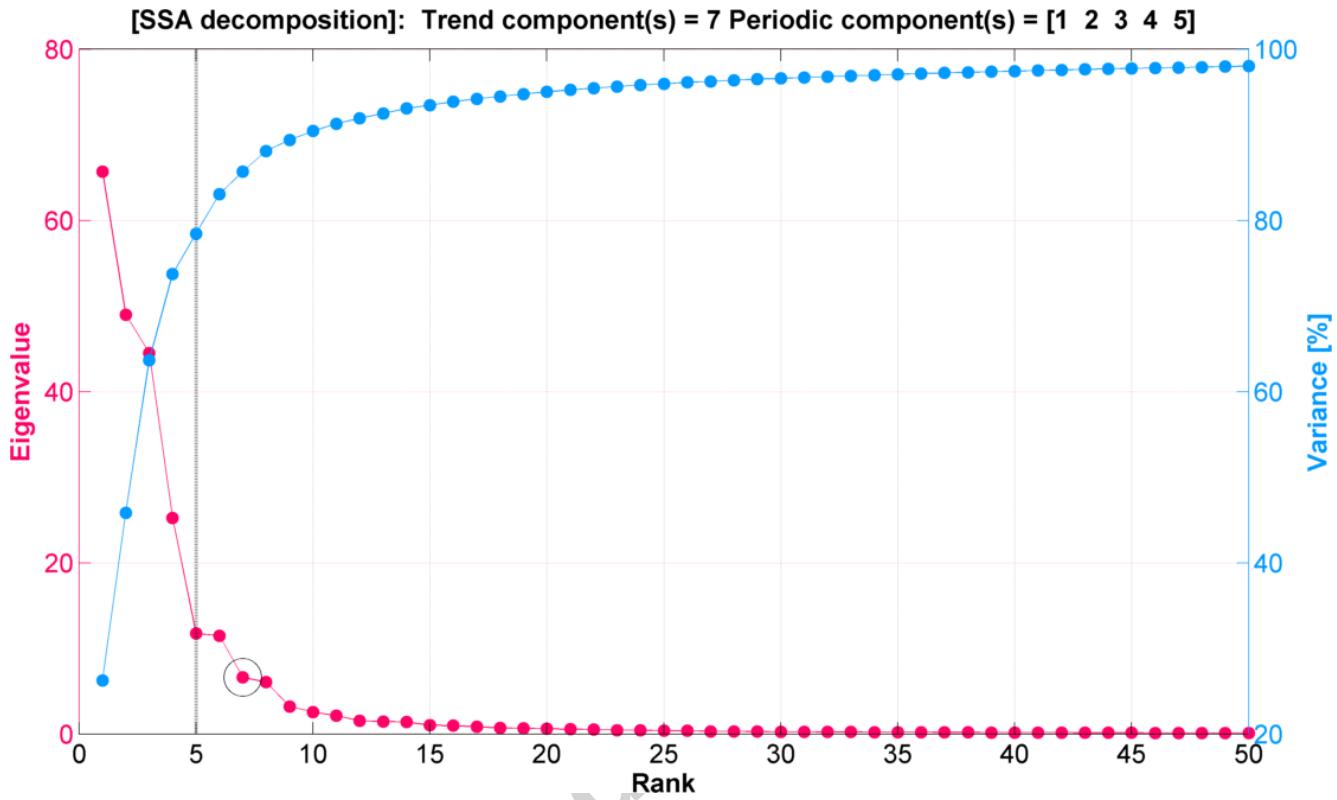


Figure 5 The rank-eigenspectrum produced by M-SSA applied to the trajectory matrix Y comprising 50 lagged time series for each of the input and output variables. The separation at the break-point between periodic components and the low plateau of the trend and noise components is shown by the vertical grey line and the trend component (eigenvalue 7) is circled on the plot.

In Fig. 6, we show the reconstructed PAR time series at Thessaloniki for 7 clear sky days during the summer of 2006 from 13-19th of August with the trend, periodic and noise components obtained with the M-SSA method and the separation point identified above. The multi-channel reconstruction error $\approx 5\%$ and the Pearson correlation of the reconstructed time series against the PAR time series is $R=0.999$.

[SSA reconstruction] Thessaloniki: 13/08/2006-19/08/2006

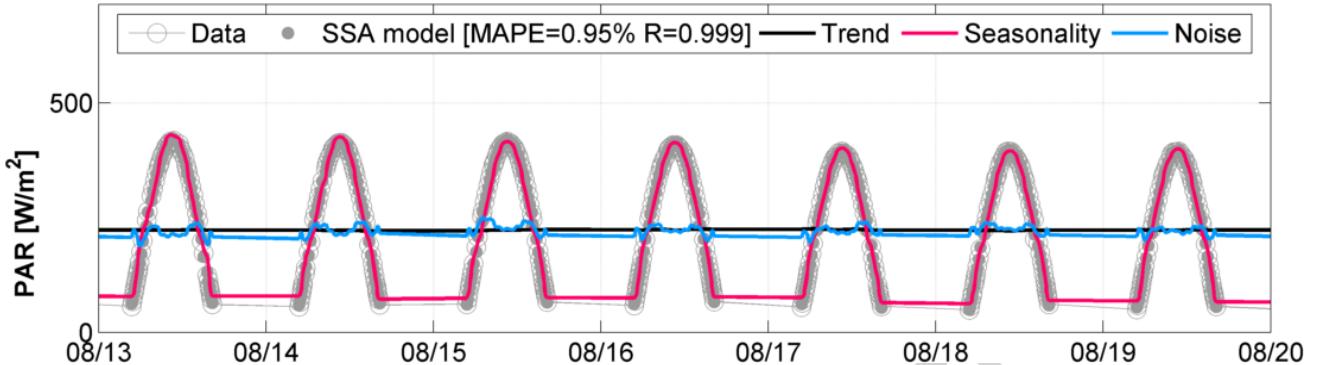


Figure 6 The PAR time series at Thessaloniki for 7 clear sky days during the summer of 2006 from the 13-19th of August with the trend, periodic and noise components obtained with the M-SSA method used in this study. The median absolute percentage error (MAPE) produced from the difference of the measured data and the reconstructed M-SSA model is also given together with the associated Pearson correlation coefficient (R).

We see that the M-SSA method appears to be an effective way to produce regularized noise models of multivariate time series data like that used in this study. The method requires 2 parameters: 1) the maximum number of time-delays (the “window length”) and 2) the number of chosen eigenvectors used for the reconstruction. Better reconstruction can be obtained by using a larger number of eigenvectors, but this has been found to be at the cost of a decrease in the smoothness of the reconstructed time series (Kandasamy et al., 2013). Returning to our choice of $M=50$ in the construction of the trajectory matrix, in the rank-eigenspectrum shown in Fig. 5, the sum of the first 8 eigenvalues accounted for 90% of the overall total of the first 50 eigenvalues, 9 eigenvalues accounted for 95%, 20 eigenvalues accounted for 99% and 74 were able to account for 99.9% of the total. In addition, we also observed that the slope of the spectrum of eigenvalues normalized to the maximum value did not change by more than $\approx 0.01\%$ for $M>50$ and so we were confident to the 99.9% level of confidence that truncating the full SVD by setting $M=50$ for each time series is a reasonable assumption. To sum up this rather technical section, M-SSA is applied as a pre-processing step to the time series of all input variables and the target irradiances to detrend and deseasonalize the data so that the remaining noise signal is regularized. The trend and periodic component is then added to the modeled irradiances to reconstruct the full model output which is denormalized and can then be validated against measurements. Before moving on to discuss the development of our NN models of the PAR and GHI irradiance in Section 4.

The performance statistics we will use to validate and test the LR, MLR and NN models are described in Appendix B.

4 Neural Network model design

4.1 NN architecture, training and validation

For multivariate input-output data like that used in this study, a feed-forward NN having at least one layer of “hidden” neurons whose activation functions are general nonlinear sigmoidal functions has been proven to have the potential to be a universal function approximator (Cybenko, 1989; Hornik, Stinchcombe and White, 1989). Given enough hidden neurons and training data, such networks are capable of learning the mathematical relation between inputs and outputs. The input vectors and output vectors used in our study as defined by Eq. 2 and Eq. 3 were connected via 2 network layers. In this study, the first layer contains 10 hidden neurons with sigmoidal activation functions, and the second contains 10 output neurons with linear activation functions. A closed form for the algebraic relation between outputs and inputs for this type of NN, which are multi-nonlinear regression models) is reported in Taylor et al. (2014). An important point to note here has been pointed out by Di Noia et al (2015) regarding the interplay between NN architectural complexity and generalization potential. An excess of neurons is known to produce over-fitting (where training errors can be small but simulation performance is poor for cases not included in the training set), but the risk of this occurring drops with the size of the training set (Di Noia et al., 2015). In this work, we follow the recommendation of Di Noia et al (2015) whereby the number of hidden layers and neurons is selected by comparing a number of different architectures and selecting the architecture that gives the lowest MSE on the validation dataset (see the discussion around Eq. 10 below). Ten hidden neurons gave consistent results for both PAR-NN and GHI-NN models. In each case, the NN model was coded using MATLAB’s object-oriented scripting language in conjunction with its Neural Network Toolbox (Beale, Hagan & Demuth, 2015). In addition to specification of the neuronal topology, a full account of NN architecture also requires a specification of how the performance error associated with the network model is to be measured, the learning procedure adopted to optimize the performance error, as well as details of any partitioning of the training dataset into “training” and “validation” components (Bishop, 1995). This is described in the next section.

4.2 NN training and validation

In our NN models, the performance error of the NN was measured via the mean squared error (MSE) calculated from the difference between NN-derived irradiances (PAR or GHI) as outputs (y_i) and the corresponding target measurements (t_i).

During each iteration of the learning process, the weights and biases are tuned to minimize the MSE cost function over the set of N input-output vectors:

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (t_i - y_i)^2 \quad (10)$$

In each training run, the number of iterations was set to 100 and training was stopped when the MSE of the validation dataset reached its minimum value. A large sample of coincident data and co-located input-output vectors for which SZA ≤ 80 and $0.2 \leq \text{PAR/GHI} \leq 0.8$ at the Thessaloniki node of HNSE, comprising 40,000 records was then split into 3 proportions such that 20,000 vectors ($\approx 42\%$) were used to train the two NN models (the “training set”), a further 10,000 ($\approx 21\%$) vectors were used to monitor the generalization ability of the NN during the training phase (the “validation set”) to ensure that no overtraining was taking place, and an equal number of 10,000 ($\approx 21\%$) independent vectors were then used to assess NN retrievals after the training phase was terminated (the “test set”). As emphasized by Di Noia et al. (2015), an important consideration for the applicability of a NN methodology to real data is that the min-max ranges and statistical properties of the training data are general enough to allow the performance of the algorithm to be valid for other realistic conditions not “seen” by the NN models during training. The range of values of each of the input and output variables used to train and test all of the models are given in Table 4.

Table 4 The min-max values, mean and standard deviation of the input and output variables used to train and test the NN, MLR and LR models on calibrated and coincident data at Thessaloniki. Note that the “training data” includes both the training set and the validation set.

Training data	N	Min	Max	Mean	St. Dev.
SZA (degrees)	30000	17.19	79.73	55.57	16.86
PAR {W/m ² }	30000	31.33	528.27	223.29	114.54
WV (cm)	30000	0.23	4.17	2.16	0.79
AOD (1)	30000	0.01	2.58	0.25	0.15
GHI (W/m ²)	30000	66.60	1170.03	504.29	249.81
Test data	N	Min	Max	Mean	St. Dev.
SZA (degrees)	10000	17.19	79.40	55.84	16.68
PAR {W/m ² }	10000	41.73	539.30	221.86	113.52
WV (cm)	10000	0.28	4.06	2.14	0.80
AOD (1)	10000	0.02	2.00	0.25	0.15

GHI (W/m ²)	10000	101.64	1189.40	501.20	247.63
-------------------------	-------	--------	---------	--------	--------

Note that data range of the variables using to test the models is well spanned by the data range of the training data. Note that the test data, while unseen by the model(s) during the training phase, at this stage is not used to test their extrapolation potential. This particular exercise is performed via the simulation runs at “external” sites reported in the results and discussion of Section 5. The NNs are trained with the Levenberg-Marquardt (LM) error backpropagation learning algorithm (Levenberg, 1944; Marquardt, 1963), which has been reported by Hagan & Menhaj (1994) to be faster and more stable than other algorithms like online backpropagation (Rumelhart et al., 1986) or conjugate gradient (Møller, 1993). While it is necessary with LM back-propagation to present the entire training dataset to the NN before each update of the network weights and that it can be slow to train the models in the case of large training sets (Wilson and Martinez, 2000), once trained, the models can simulate without further calculation (Di Noia et al., 2015). In Fig. 7 the course of the NN learning is presented for the PAR-NN as for the case of data with M-SSA pre-processing (the GHI-NN curve is similar). Pre-processing of the input-output time series with M-SSA has the desirable effect that the descent of the learning curve is smooth. Note also that the models regularized with SSA have undergone normalization and this is why the validation MSE is low in Figure 7. The performance statistics for the training, validation and testing of the PAR-NN and GHI-NN regularized by M-SSA are summarized in Table 5.

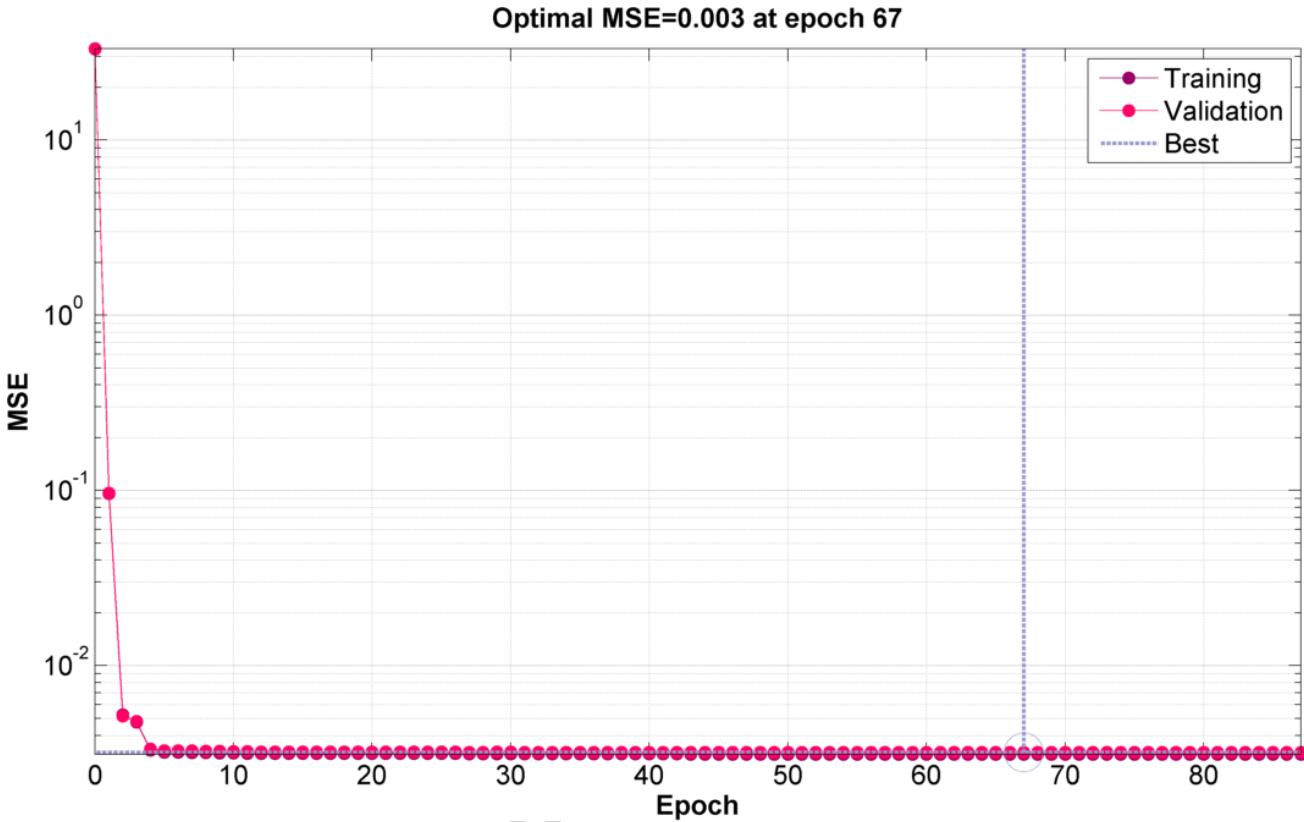


Figure 7 Network training plot showing the variation of the MSE (as defined by Eq. 10) between target values and NN outputs with epoch of the learning process for the PAR-NN with M-SSA.

Table 5 NN training performance statistics (described in Section 3) for the retrieved GHI and PAR irradiances (y) and target measurements (t) computed for both the PAR-NN and GHI-NN models on the training, test and validation set. The statistics are separated into exploratory, 1st moment (linear) and 2nd moment (quadratic) categories.

Statistics	PAR-NN	PAR-NN	PAR-NN	GHI-NN	GHI-NN	GHI-NN
	training	validation	Test	training	validation	test
N	20000	10000	10000	20000	10000	10000
mean(t)	222.725	224.405	221.862	503.075	506.722	501.198
std(t)	114.492	114.641	113.522	249.665	250.105	247.627
mean(y)	222.729	224.429	221.816	503.057	506.591	501.127
std(y)	114.049	114.188	113.068	249.423	249.788	247.254
b	0.004	0.024	-0.046	-0.017	-0.131	-0.071
MAE	4.599	4.637	4.651	10.151	10.268	10.297
r	0.998	0.998	0.998	0.998	0.998	0.998
d ₁	0.977	0.977	0.977	0.977	0.977	0.976

RMSE	6.375	6.461	6.741	14.000	14.384	14.814
RMSE(s)	0.620	0.634	0.655	0.634	0.742	0.819
RMSE(u)	6.345	6.430	6.709	13.986	14.365	14.792
R ²	0.997	0.997	0.996	0.997	0.997	0.996
d ₂	0.999	0.999	0.999	0.999	0.999	0.999

The training performance statistics suggest that there is consistency between the NN irradiance outputs for the training, validation and test sets. It is important that for both models, the measure of model precision expressed via RMSE_u for the validation and test sets < 3% of that obtained for the training set and never exceeds 7 W/m² for PAR and 15 W/m² for GHI. This is echoed also by the size of the average error for all 3 data sets. The size of the mean bias is also low (< 0.14W/m²), and the models appear to capture extremely well the variance in the data as expressed by the range of values of the coefficient of determination 0.996 ≤ R² ≤ 0.997. This is also reflected visually in the reconstructed PAR-NN output for the test set shown in Fig. 8.

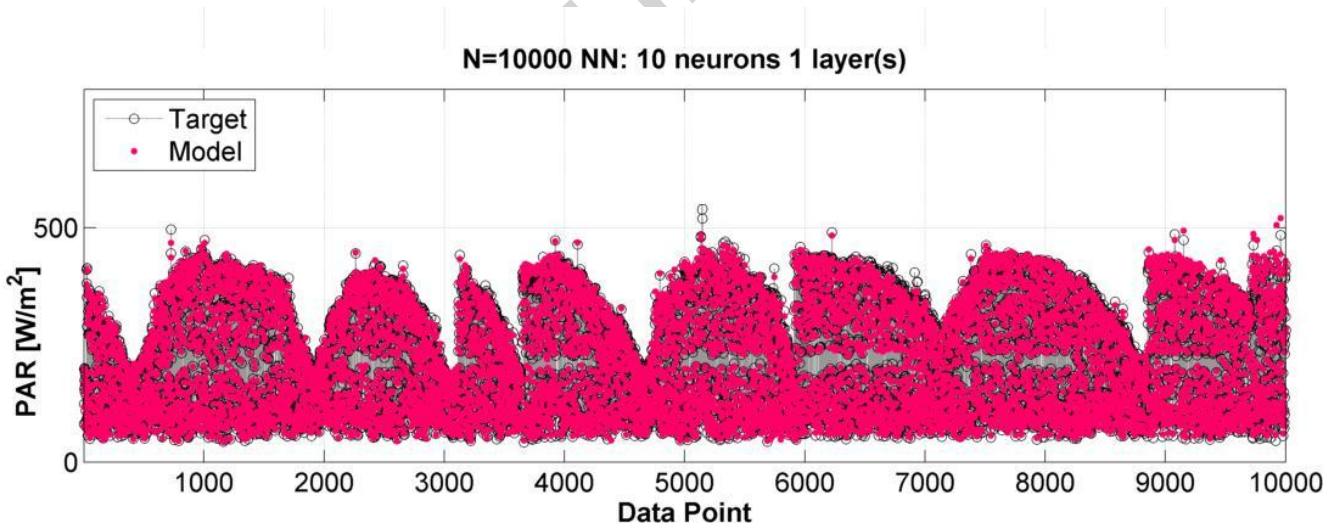


Figure 8 Target measurements in the test set (grey circles) and the outputs from the PAR -NN model (red dots).

4.3 Linear versus nonlinear models

At this point, we assess the performance of modelling PAR and GHI with simpler linear regression models having the general multivariate form:

$$\text{PAR}_{\text{MLR}} = b_1 \times \text{SZA} + b_2 \times \text{GHI} + b_3 \times \text{WV} + b_4 \times \text{AOD} + c_2 \quad (11)$$

$$GHI_{MLR} = a_1 \times SZA + a_2 \times PAR + a_3 \times WV + a_4 \times AOD + c_1 \quad (12)$$

LR models of GHI and PAR correspond to the case $a_1=a_3=a_4=b_1=b_3=b_4=0$. For the same 30,000 coincident records of combined training and validation data at Thessaloniki used to train and validate the NN models, regression of the responses (the LHS of Eq. 11 and Eq. 12) on the matrix of predictors formed from the input variables (the corresponding RHS of Eq. 11 and Eq. 12), gave the following coefficient estimates for the linear regression (LR) and multiple linear regression (MLR) models of the target PAR and GHI irradiances with coefficients estimated via least squares optimization at the 95% level of confidence (upper and lower bounds on the coefficients are not shown for clarity):

$$PAR_{LR} = 0.457 GHI - 7.424 \quad (13)$$

$$PAR_{MLR} = -0.375 SZA + 0.431 GHI + 3.673 WV - 3.447 AOD + 19.940 \quad (14)$$

$$GHI_{LR} = 2.176 PAR + 18.398 \quad (15)$$

$$GHI_{MLR} = -1.439 SZA + 1.983 PAR + 10.769 WV - 7.555 AOD + 166.615 \quad (16)$$

Table 6 presents the training performance statistics for the LR and MLR models of PAR (given by Eq. 13 and Eq. 14 respectively). Similar performance was found for LR and MLR models of GHI.

Table 6 Training performance statistics for the retrieved PAR irradiances (y) and target measurements (t) computed with LR and MLR models (data sets and statistics used are the same as in Table 5)

Statistics	LR	LR	LR	MLR	MLR	MLR
	training	validation	test	training	validation	test
N	20000	10000	10000	20000	10000	10000
mean(t)	222.725	224.405	221.862	222.725	224.405	221.862
std(t)	114.492	114.641	113.522	114.492	114.641	113.522
mean(y)	222.715	224.426	221.811	222.720	224.416	221.811
std(y)	114.196	114.379	113.237	114.241	114.403	113.270
b	-0.010	0.021	-0.052	-0.006	0.011	-0.051
MAE	4.961	4.952	5.010	4.755	4.768	4.789
r	0.998	0.998	0.998	0.998	0.998	0.998
d ₁	0.975	0.976	0.975	0.976	0.976	0.976
RMSE	6.813	6.863	7.128	6.613	6.663	6.886
RMSE(s)	0.499	0.468	0.511	0.442	0.431	0.463
RMSE(u)	6.794	6.847	7.109	6.598	6.649	6.871

R^2	0.996	0.996	0.996	0.997	0.997	0.996
d_2	0.999	0.999	0.999	0.999	0.999	0.999

The training performance statistics of the LR and MLR models of the PAR above are comparable with those for the NN model of PAR presented in Table 4 so that for example, $r=0.999$ in all 3 models, the mean bias is very low, close to zero and marginally negative in the range: $-0.052 \leq b \leq -0.046 \text{ W/m}^2$, and the average errors are in the range: $4.651 \leq \text{MAE} \leq 5.010 \text{ W/m}^2$ or $6.709 \leq \text{RMSE}(u) \leq 7.109 \text{ W/m}^2$ for the test data. The NN model slightly out-performs the MLR and LR models here. We would like to say that the fact that the noise component only represents a small fraction of the time series' variance, means that the impact of the SSA-processed LR, MLR and NN models is swamped by the variance contained in the signal (combined trend and periodicity). However, for studies of other geophysical quantities where the signal to noise ratio is very different (e.g. 1:1) then the quality of the model of the noise is more important. In such cases, it is expected that NN models will greatly outperform their linear counterparts. In our study, it may be expected then, that all models will also perform more or less equally well in unsupervised mode where they are used for simulations with satellite-corrected inputs (described in Section 3.2). This comparison is made in order to answer the question of whether or not there is any advantage to be gained (in terms of model performance) by developing a nonlinear NN model instead of a linear or multilinear regression model of the same complexity (i.e. choice of input parameters). It is important to recall here that in Table 5 and Table 6, training performance statistics are associated with time series variables that have been pre-processed and reconstructed with M-SSA. As such, the outputs from the LR, MLR and NN models are reconstructed from the additive sum of the “stationary” signal (the extracted trend and periodic components) and the modeled noise component. The very high signal to noise ratio arising from the presence of strong periodicity (see the discussion related to the diurnal cycle and other deterministic effects in Section 3.3) means that the reconstructed model outputs are expected to be dominated by the stationary signal rather than the modeled noise. In the next section it will become clear that LR, MLR and NN models that are regularized with M-SSA perform generically and are quite accurate (with some small advantage in the case of the NN models) due to the ability of M-SSA to capture $\approx 95\%$ of the variance of the time series. We will see that it is this feature of M-SSA that gives all 3 models their high level of simulation performance. It is also the reason why all 3 models perform more poorly on days where the

noise component is large (i.e. the signal to noise ratio is lower). On these days, proper regularization of the time series with M-SSA is paramount. We summarize the rational for the modeling approach in Fig. 9.

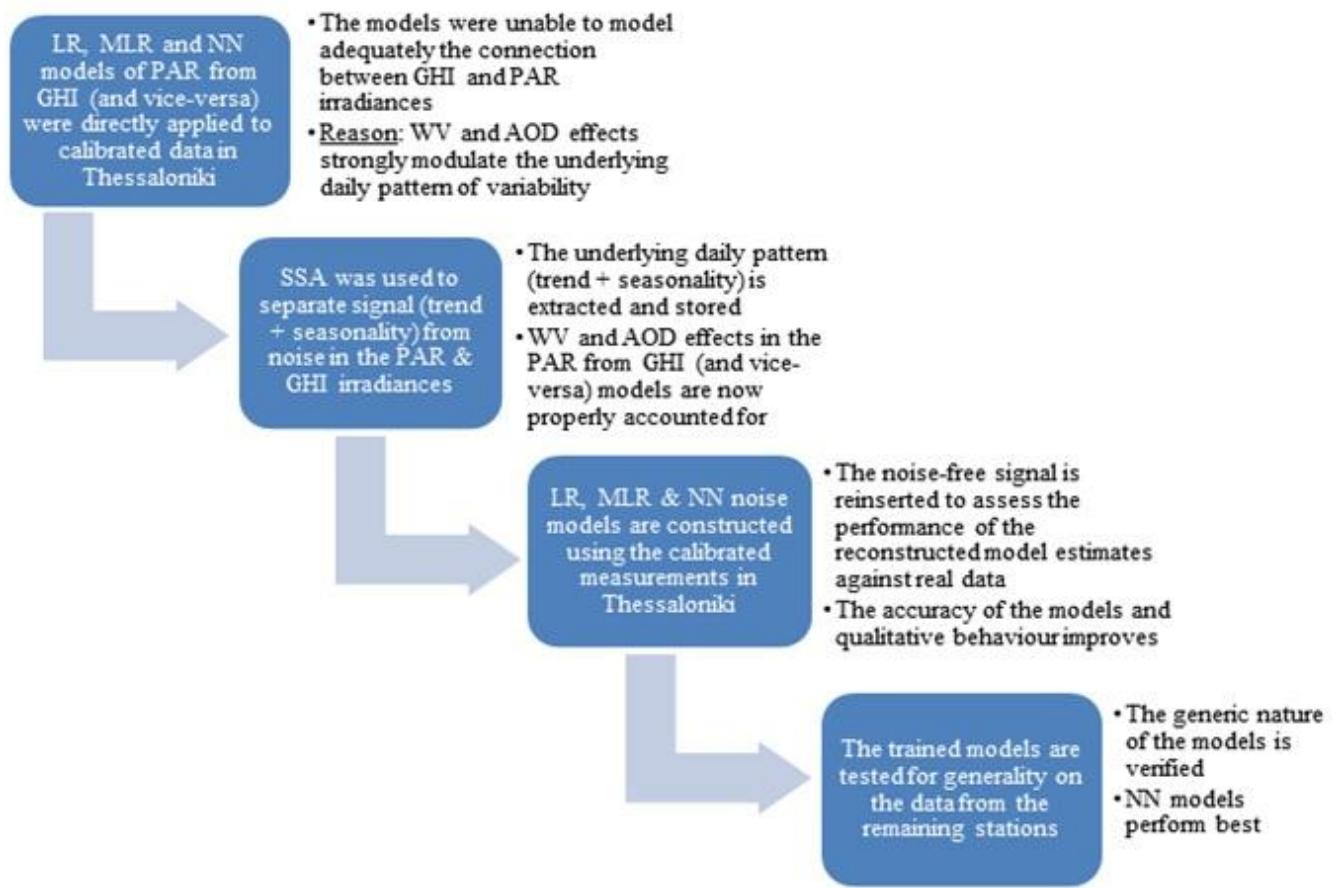


Figure 9 A schematic of the rationale for the model configuration and their implementation.

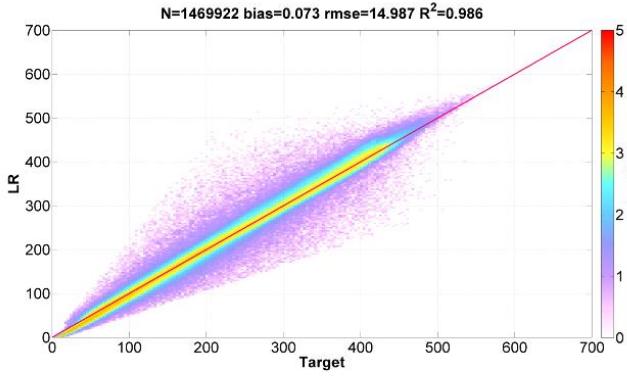
5. Results and discussion

We now turn to a presentation of the results of running trained models (LR, MLR and NN) of PAR and GHI with the satellite-corrected inputs described in Section 3.2. We will present model simulations at the training site in Thessaloniki and also at the “external” sites (ATH, FIN, IOA, MYT, XAN) so as to test the generic nature and spatial extrapolation capability and applicability of the trained models.

5.1 The impact of M-SSA on LR, MLR and NN simulations

In Fig. 10, log density scatterplots of the PAR outputs of the LR, MLR and NN models regressed on coincident satellite-corrected PAR measurements at the Thessaloniki site are presented for the case of data that is regularized and reconstructed with or without M-SSA. Without M-SSA, the PAR-MLR and especially the PAR-NN models show a large degradation in performance at lower values of measured PAR where large dispersion is apparent. This is reflected in a decline in the Pearson correlation coefficient R (e.g. from 0.988 to 0.949 in the case of the GHI-NN model), a substantial increase in the RMSE (e.g. from 30.714 to 59.741 in the case of the PAR-NN model) and the appearance of a substantial mean bias, b (e.g. from -0.020 to -6.728 in the case of the PAR-MLR model). The PAR-LR model appears to be more robust.

Models not regularized with M-SSA



Models regularized with M-SSA

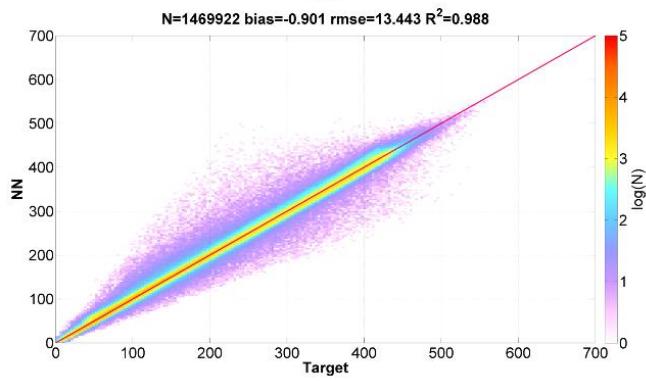
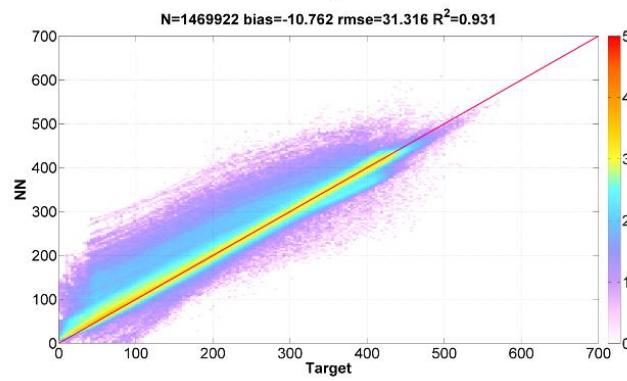
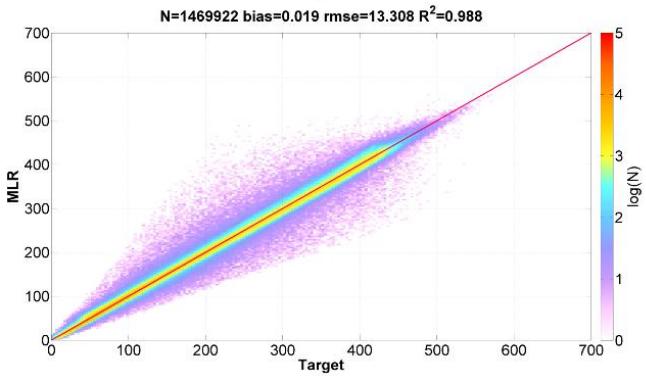
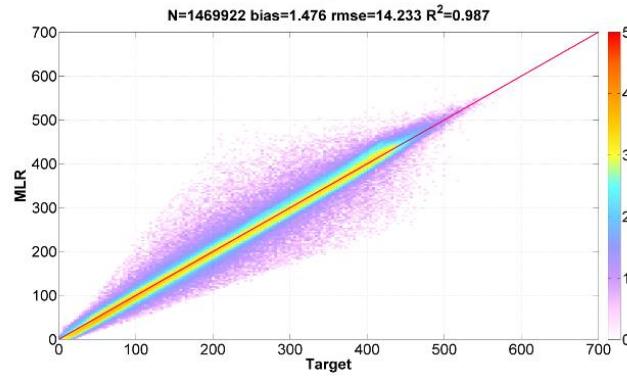
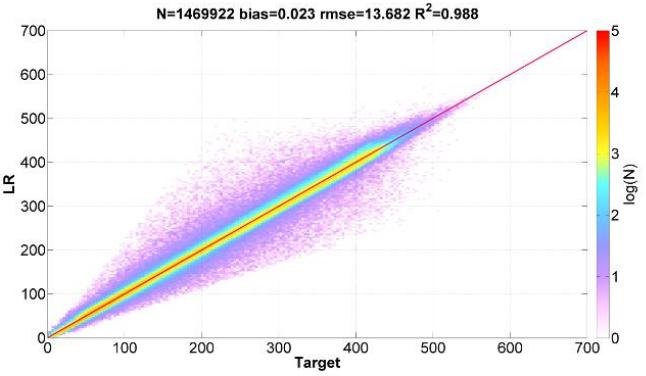


Figure 10 Log density scatterplots of the outputs of the LR, MLR and NN models regressed on coincident satellite-corrected PAR measurements at the Thessaloniki site in the HNSE. The left hand panels are for models trained on time series regularized and reconstructed with M-SSA. The right hand panels are without M-SSA processing.

However, inspection of the variation of the modeled time series on the daily timescale shows that regularization with M-SSA is very important for constraining all 3 models as can be seen in Fig. 11. A number of effects are visible. First of all, in the upper panels of Fig. 11, the model simulations appear to diverge from the data after the peak of the irradiance. This is due to

the arrival of clouds in the scene (visible as “absorption troughs”) which disrupt the background “clear sky” irradiance curve. While M-SSA is not able to fully correct for this effect, a minor correction is apparent despite the sporadic nature of clouds. Such corrections are more stark in the regularization of the scattered cloud and overcast irradiance curves of the middle and lower panels of Fig. 11 given by M-SSA.

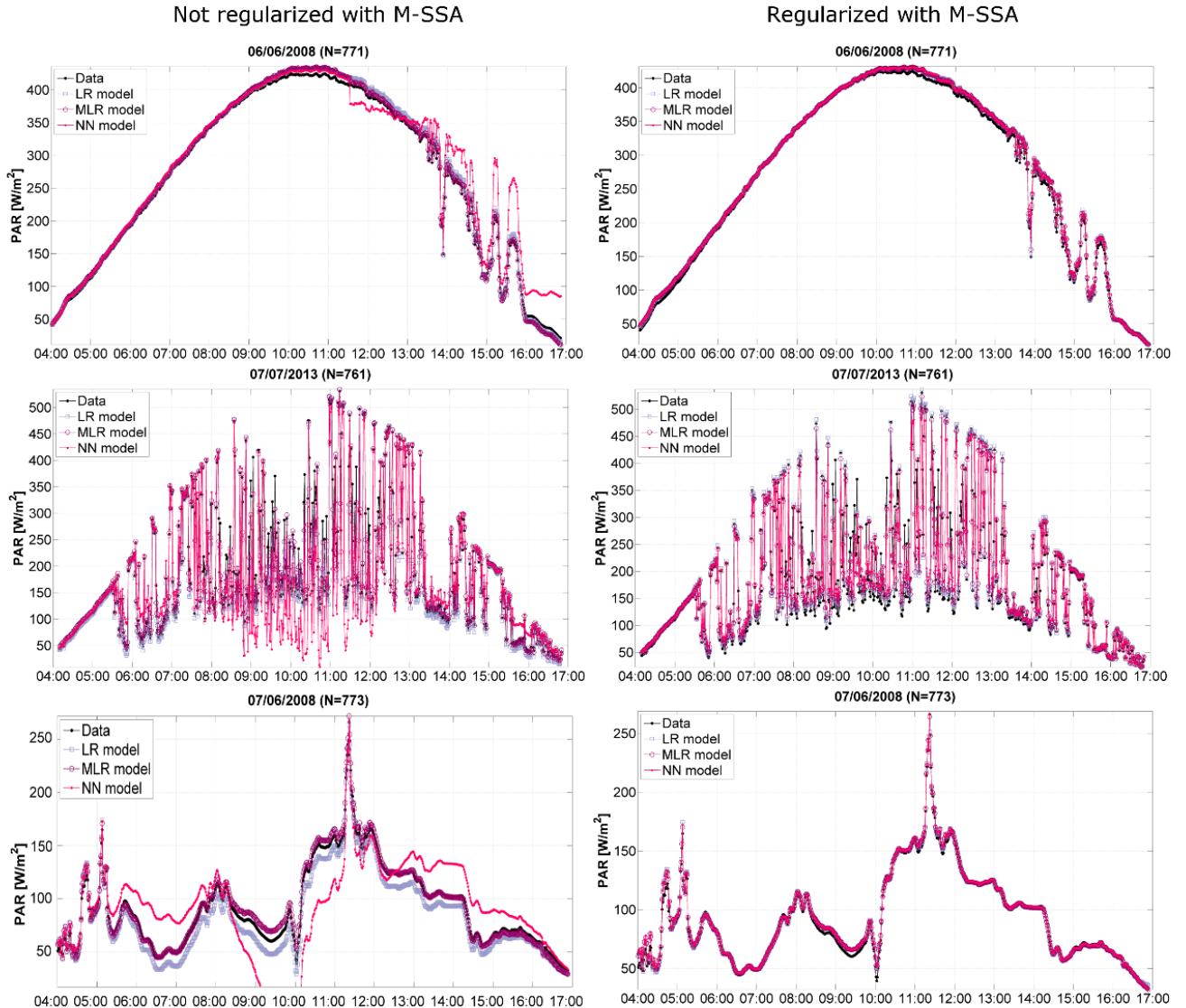


Figure 11 The fit of the outputs of the LR, MLR and NN models to coincident PAR measurements during a clear sky day (top panels), a day with scattered clouds (middle panels) and overcast conditions (lower panels). The left hand panels are for models trained on time series regularized and reconstructed with M-SSA. The right hand panels are without SSA processing.

In order to assess the quality of the M-SSA reconstruction, Table 7 summarizes the variance accounted for by the extracted trend, periodic and noise components at the training site and the external sites used to assess the simulation performance of the LR, MLR and NN models of PAR and GHI.

Table 7 Statistics of the M-SSA reconstruction of the modeled input-output time series at the “training site” (Thessaloniki) and the 5 “external sites” (Athens, Finokalia, Ioannina, Mytilene and Xanthi).

M-SSA	THE	ATH	FIN	IOA	MYT	XAN
Records (N)	1469922	188850	390956	309922	417755	295514
Variance of Trend (%)	0.76	1.20	1.01	1.45	1.04	1.40
Variance of Periodic (%)	95.41	94.91	95.62	93.73	95.90	94.18
Variance of Noise (%)	3.27	3.27	2.70	4.10	2.48	3.74
Variance unaccounted for (%)	0.56	0.62	0.68	0.72	0.58	0.67

As expected from the discussion of the signal to noise ratio in Section 4.3, it is clear that the most dominant component of the M-SSA time series decomposition is the periodic component which accounts for between 94.18% and 95.90% of the total variance in the input-output time series at all sites used in this study. It is noteworthy also that the noise component (ranging from 2.48% to 3.74%) accounts for much more of the variance than the trend component (which ranges from 0.76% to 1.45%). For all sites, the M-SSA reconstruction accounts for at least 99.28% of the total variance (even though this is expected to vary on individual days). The unaccounted-for variance arises from the SVD truncation error associated with performing the reconstruction with M=50 eigenvectors as described in Section 3.3.

In Table 8 we present the MAPE arising from a comparison of the satellite-corrected PAR and GHI target data with the M-SSA reconstruction and also the reconstruction of only the signal (without the noise component). The purpose of this exercise is to assess the impact of the noise component in our LR, MLR and NN models on the fit to the target data. For both PAR and GHI, the reconstructed model outputs, correlate strongly with the target satellite-corrected data ($0.998 \leq r \leq 0.999$), and the error of the reconstruction (the “SSA error”) is very low ($0.16\% \leq \text{MAPE} \leq 0.62\%$) despite the fact that there is an intrinsic error due to unaccounted for variance in the truncated SVD. However, when we compare the target PAR and GHI with the reconstructed trend and periodic component then we see that the highest reconstruction errors are at the IOA and XAN sites, and correspond to high MAPEs for the reconstructed trend and periodic components (i.e. the “stationary” signal).

The corollary of this is that these errors are due to the presence of higher variance noise components at these 2 sites. A look at Table 8 confirms that this is the case. We therefore expect that the LR, MLR and NN models of PAR and GHI without M-SSA pre-processing to be error prone.

Table 8 Statistics of the data versus time series reconstruction of PAR and GHI at the “training site” (Thessaloniki) and the 5 “external sites” (Athens, Finokalia, Ioannina, Mytilene and Xanthi) resulting from the M-SSA of **Table 7**.

PAR	THE	ATH	FIN	IOA	MYT	XAN
MAPE: SSA stationary signal (%)	2.47	1.94	2.94	4.41	1.74	3.15
MAPE: SSA reconstruction (%)	0.30	0.22	0.41	0.55	0.16	0.42
Pearson r : SSA stationary signal	0.980	0.970	0.977	0.959	0.978	0.964
Pearson r : SSA reconstruction	0.998	0.999	0.998	0.998	0.999	0.998
GHI	THE	ATH	FIN	IOA	MYT	XAN
MAPE: SSA stationary signal (%)	2.43	1.99	2.99	4.94	1.73	3.26
MAPE: SSA reconstruction (%)	0.31	0.23	0.47	0.62	0.21	0.45
Pearson r : SSA stationary signal	0.978	0.968	0.975	0.957	0.976	0.962
Pearson r : SSA reconstruction	0.998	0.999	0.998	0.998	0.999	0.998

The M-SSA based models developed here therefore, produce valid results for one of the two radiometric values when only the second is available (in conjunction with supplementary geolocation and aerosol data that can be provided by satellite instruments). In particular, since GHI measurements are generally more common, these can be used to derive PAR reconstructed data of high accuracy at sites around the globe. Spatial interpolation methods can then be used to enrich the spatial distribution of both quantities worldwide. Also, in the reverse situation such as the measurements from the Greek UVNET where PAR measurements are available since 2005 but GHI measurements are of more limited duration (HNSE, since 2011), the GHI can be reconstructed from the NN model to augment the time series of available measurements, and thus result in a longer accurate dataset for statistical analysis of GHI levels. In addition to the production of regularized noise models like the ones used in this study, SSA has also been shown to enable powerful time series forecasting capability (see for example Hassani et al., 2009 and Golyandina & Korobeynikov, 2014). Key to this is that SSA can provide a basis for the construction of forecast models by forward extrapolation of trend and periodic components in conjunction with either a linear ARMAX model or a nonlinear NARX-NN model of the modeled noise component using lag values. The advantage of

SSA applied in this way is that the removal of trend and periodic components as a pre-processing step and their reintegration into model output as a post-processing step, means that the forecast model is reduced to the problem of forecasting only the regularized noise component, and is therefore better constrained.

5.2 PAR simulations

When summarizing the simulation performance statistics of the difference between the outputs of the LR, MLR and NN models of the GHI and the target measurements at the training site and the 5 external sites when M-SSA processing is applied, we note that there is a moderate degree of variation with site as can be seen from the distribution of differences in Fig. 12 for the NN model of PAR. The mean (μ) of the differences at each site occupies a small range ($0.859 \leq \mu \leq 1.50$) and the standard deviation of the differences is relatively broad ($12.12 \leq \sigma \leq 23.87$). In Table 9, we present the simulation performance statistics for the PAR-NN model with site.

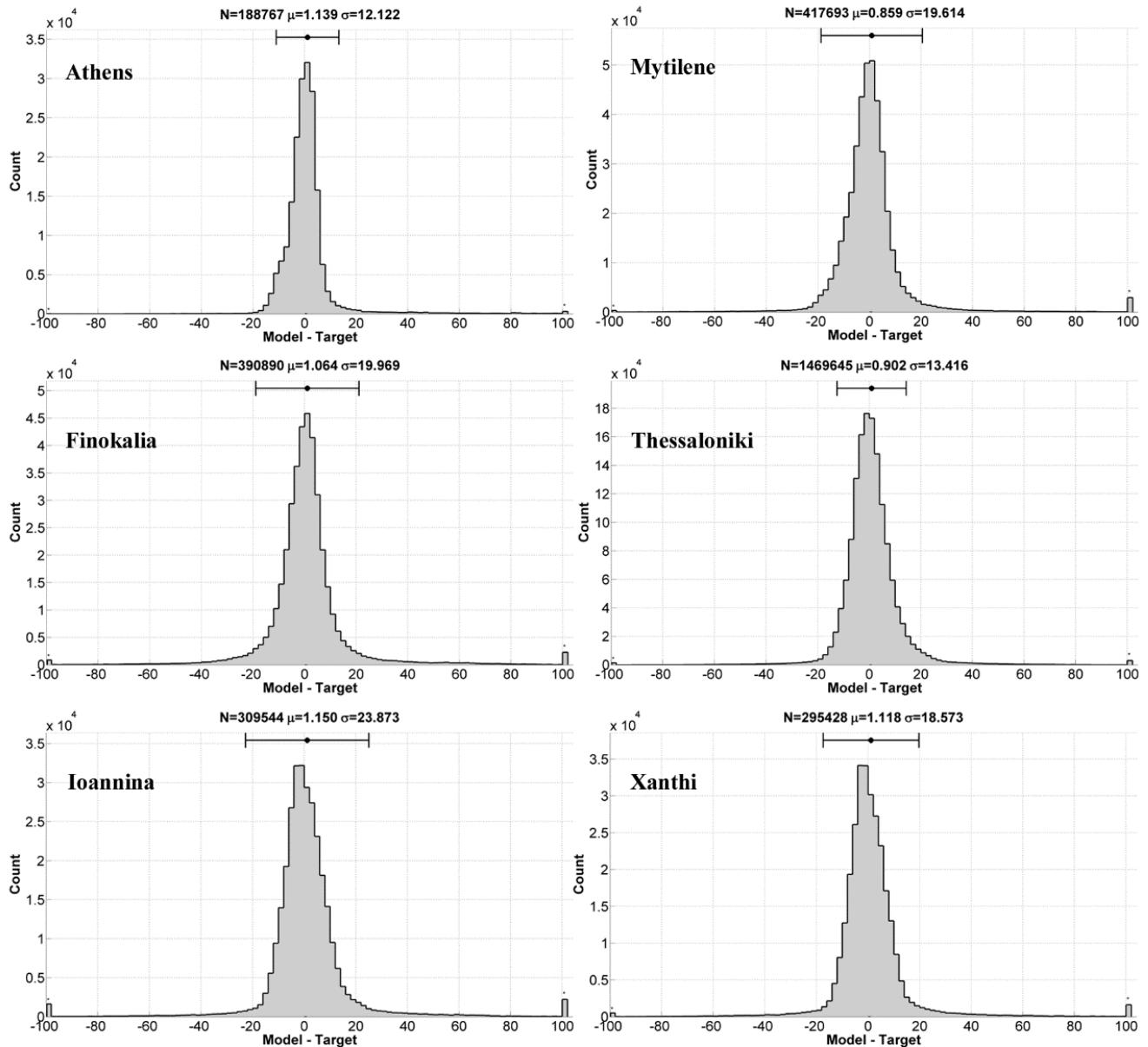


Figure 12 Histograms of the difference between the outputs of the PAR-NN and the time series of the target measurements at the “training site” (Thessaloniki) and the 5 “external sites” (Athens, Finokalia, Ioannina, Mytilene and Xanthi). In all cases, the number of bins is equal to 200, the best fit Gaussian distribution with mean difference (μ) and standard deviation (σ) is shown.

Table 9 Statistics of the difference between the outputs of the PAR-NN and the time series of the target measurements at the “training site” (Thessaloniki) and the 5 “external sites” (Athens, Finokalia, Ioannina, Mytilene and Xanthi) with M-SSA processing.

PAR-NN	THE	ATH	FIN	IOA	MYT	XAN
N	1469645	188767	390890	309544	417693	295428

mean(t)	210.117	223.382	221.208	181.774	223.777	202.657
std(t)	123.150	125.664	134.848	120.490	118.421	123.105
mean(y)	211.019	224.521	222.272	182.924	224.636	203.775
std(y)	122.806	124.839	134.274	119.586	117.810	122.319
b	0.902	1.139	1.064	1.150	0.859	1.118
MAE	7.217	5.780	9.618	10.368	8.545	8.706
r	0.994	0.995	0.989	0.980	0.986	0.989
d ₁	0.966	0.973	0.959	0.949	0.958	0.959
RMSE	13.446	12.176	19.998	23.901	19.633	18.606
RMSE(s)	1.403	1.811	2.311	3.462	2.393	2.455
RMSE(u)	13.373	12.040	19.864	23.649	19.487	18.444
R ²	0.988	0.991	0.978	0.961	0.973	0.977
d ₂	0.997	0.998	0.994	0.990	0.993	0.994

The LR, MLR and NN models of PAR largely agree with one another as shown by the min-max of the statistics shown in Table 10.

Table 10 Summary of the simulation performance statistics of LR, MLR and NN models of the PAR with M-SSA processing across the training and external sites.

	b	MAE	r	d ₁	RMSE	RMSE(s)	RMSE(u)	R ²	d ₂
Min (NN)	0.859	5.780	0.980	0.949	12.176	1.403	12.040	0.961	0.990
Max (NN)	1.150	10.368	0.995	0.973	23.901	3.462	23.649	0.991	0.998
Mean (NN)	1.039	8.372	0.989	0.961	17.960	2.306	17.809	0.978	0.994
St. Dev. (NN)	0.127	1.657	0.005	0.008	4.393	0.697	4.347	0.011	0.003
Min (MLR)	-0.020	5.313	0.980	0.950	11.186	0.908	11.148	0.960	0.990
Max (MLR)	-0.012	10.177	0.996	0.976	24.149	2.988	23.963	0.992	0.998
Mean (MLR)	-0.017	8.169	0.989	0.962	17.785	1.729	17.698	0.978	0.994
St. Dev. (MLR)	0.003	1.745	0.006	0.009	4.744	0.769	4.694	0.012	0.003
Min (LR)	-0.024	5.263	0.979	0.949	11.373	0.445	11.364	0.958	0.989
Max (LR)	-0.015	10.443	0.996	0.976	24.790	2.590	24.654	0.992	0.998
Mean (LR)	-0.021	8.313	0.988	0.961	18.233	1.376	18.177	0.977	0.994
St. Dev. (LR)	0.003	1.834	0.006	0.009	4.895	0.772	4.852	0.012	0.003

There is also close agreement in the average simulation performance of all 3 models across the training and external sites. In terms of linear statistics, the mean bias is again low, close to zero and in the range: $-0.017 \leq b \leq 1.039 \text{ W/m}^2$, the MAE is as expected, lower than in the case of the GHI, and is in the range: $8.169 \leq \text{MAE} \leq 8.372 \text{ W/m}^2$, r is in the range $0.988 \leq r \leq 0.989$, and d_1 is in the range $0.961 \leq d_1 \leq 0.962$. The second order statistics report minor differences between the average performance of the models - with the NN and MLR models performing marginally better than the LR model on average across all sites.

5.3 GHI simulations

Summarizing the simulation performance statistics of the difference between the outputs of the LR, MLR and NN models of the PAR and target measurements at the training and external sites, just as PAR, we note that there is a moderate degree of variation with site as can be seen from the distribution of differences in Fig. 13 for the GHI-NN model. While the mean (μ) of the differences at each site occupies, once again, a small range ($-0.28 \leq \mu \leq -0.78$), the standard deviation of the differences is much broader ($26.12 \leq \sigma \leq 56.48$). In Table 11, we present the simulation performance statistics for the NN model with site.

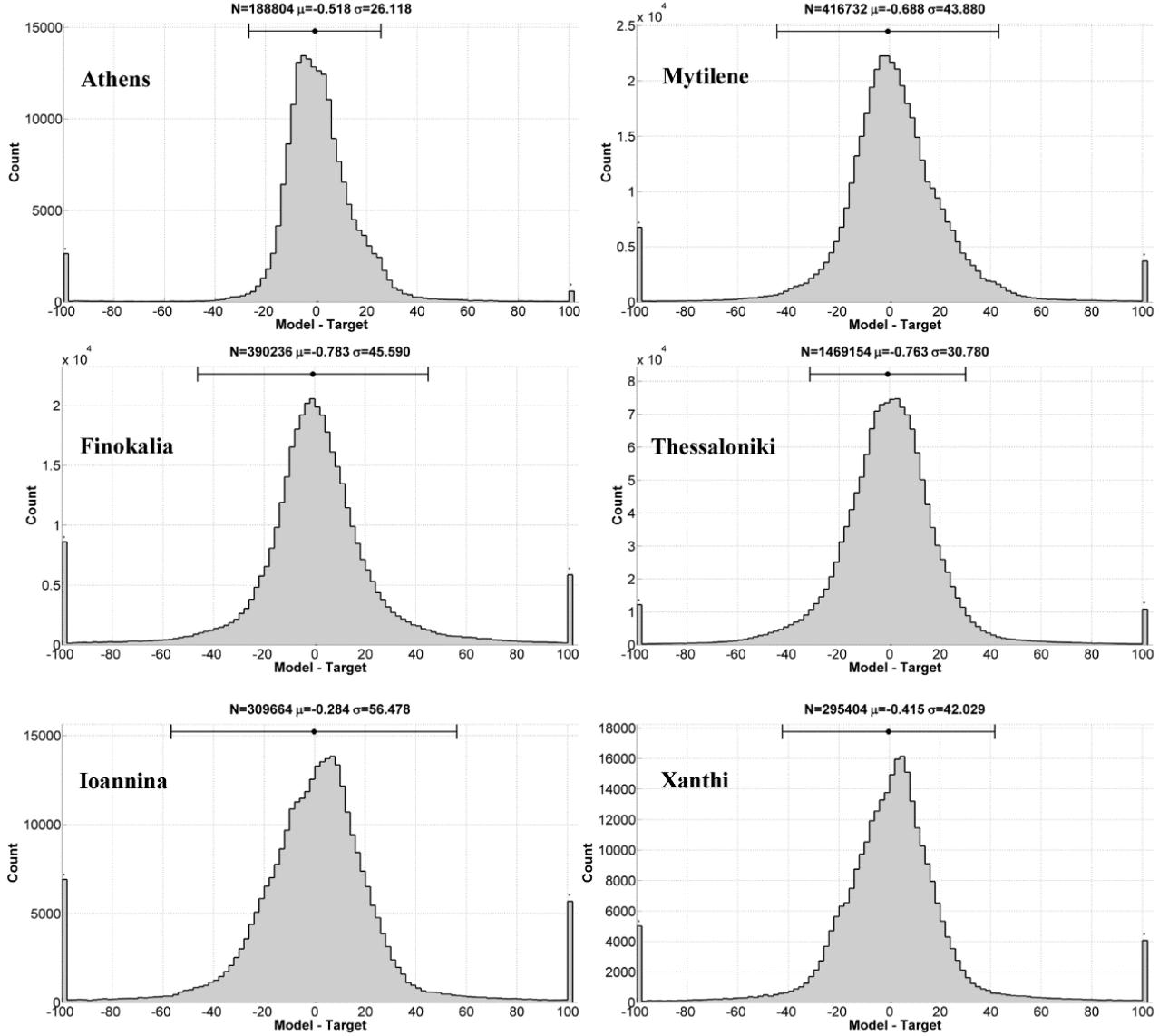


Figure 13 Histograms of the difference between the outputs of the GHI-NN and the time series of the target measurements at the “training site” (Thessaloniki) and the 5 “external sites” (Athens, Finokalia, Ioannina, Mytilene and Xanthi). In all cases, the number of bins is equal to 200, the best fit Gaussian distribution with mean difference (μ) and standard deviation (σ) is shown.

Table 11 Statistics of the difference between the outputs of the GHI-NN model and the time series of the target measurements at the “training site” (Thessaloniki) and the 5 “external sites” (Athens, Finokalia, Ioannina, Mytilene and Xanthi) with M-SSA processing.

GHI-NN	THE	ATH	FIN	IOA	MYT	XAN
N	1469154	188804	390236	309664	416732	295404

mean(t)	475.504	500.009	498.409	422.305	511.020	455.307
std(t)	276.568	280.290	305.976	279.678	269.319	277.698
mean(y)	474.741	499.491	497.626	422.021	510.332	454.892
std(y)	275.315	278.344	304.126	278.187	267.768	276.026
b	-0.763	-0.518	-0.783	-0.284	-0.688	-0.415
MAE	16.702	12.822	21.799	24.579	19.269	19.861
r	0.994	0.996	0.989	0.980	0.987	0.988
d ₁	0.965	0.974	0.959	0.948	0.959	0.959
RMSE	30.790	26.123	45.596	56.479	43.885	42.031
RMSE(s)	3.060	3.199	5.299	7.195	5.167	4.865
RMSE(u)	30.637	25.926	45.287	56.019	43.580	41.749
R ²	0.988	0.991	0.978	0.959	0.973	0.977
d ₂	0.997	0.998	0.994	0.990	0.993	0.994

As in PAR simulations, despite the existence of a small variation of simulation performance with site, the 3 models however, largely agree with one another as shown by the min-max of the statistics shown in Table 12.

Table 12 Summary of the simulation performance statistics of LR, MLR and NN models of the GHI with M-SSA processing across the training and external sites.

	b	MAE	r	d₁	RMSE	RMSE(s)	RMSE(u)	R²	d₂
Min (NN)	-0.783	12.822	0.980	0.948	26.123	3.060	25.926	0.959	0.990
Max (NN)	-0.284	24.579	0.996	0.974	56.479	7.195	56.019	0.991	0.998
Mean (NN)	-0.575	19.172	0.989	0.961	40.817	4.798	40.533	0.978	0.994
St. Dev. (NN)	0.202	4.075	0.006	0.009	10.913	1.531	10.812	0.011	0.003
Min (MLR)	0.019	12.615	0.979	0.948	25.766	3.103	25.567	0.959	0.990
Max (MLR)	0.090	24.556	0.996	0.974	56.485	7.810	55.942	0.992	0.998
Mean (MLR)	0.036	19.146	0.989	0.961	40.999	4.970	40.694	0.978	0.994
St. Dev. (MLR)	0.028	4.149	0.006	0.009	11.127	1.726	11.005	0.012	0.003
Min (LR)	0.048	11.996	0.978	0.948	25.770	1.861	25.702	0.957	0.989
Max (LR)	0.125	24.651	0.996	0.975	58.255	6.685	57.870	0.992	0.998
Mean (LR)	0.067	19.202	0.988	0.961	42.040	3.991	41.846	0.976	0.994
St. Dev. (LR)	0.031	4.419	0.006	0.009	11.766	1.720	11.657	0.012	0.003

There is close agreement in the average simulation performance of all 3 models across the training and external sites. In terms of linear statistics across sites and for each of the 3 models, the mean bias is low, close to zero and in the range: $-0.575 \leq b \leq 0.067 \text{ W/m}^2$, the MAE is in the range: $19.146 \leq \text{MAE} \leq 19.202 \text{ W/m}^2$, r is in the range $0.988 \leq r \leq 0.989$, and $d_i=0.961$ to 3 decimal places. Second order statistics which are more sensitive to the size of differences suggest that the NN model is, on average, slightly more precise with RMSE (u)=40.533 than the MLR and LR models where RMSE(u)= 40.694 and 41.846 W/m^2 respectively. This is also reflected in a slightly higher average value of $R^2=0.978$ for the NN model as compared with $R^2=0.976$ for the LR model. While these are rather minor differences, they suggest that the nonlinear NN model consistently performs better than its linear counterparts (LR and MLR).

5.4 Generality and applicability

The results of Section 5.1 and Section 5.2 demonstrate that the linear and nonlinear models, despite being trained on data at a single site, achieve excellent performance at all sites and thus exhibit generic capabilities. M-SSA has demonstrated a capacity for i) regularizing models constructed from an array of highly variable time series and ii) gives extrapolation power to both linear regression (LR and MLR) and nonlinear multivariate models (NN) of PAR and GHI irradiances. This study therefore suggests that LR, MLR and NN models of PAR and GHI are generic provided that M-SSA is applied to input-output time series.

6. Conclusions

In this study, robust linear regression and nonlinear NN time series models that implicitly include the effect of cloud and aerosol physics, are introduced in order to access PAR from GHI measurements (and vice versa). The models were trained, validated and tested on calibrated and high temporal resolution measurements taken at the station of Thessaloniki where both ground-based measurements of PAR and GHI as well as satellite AOD and WV data and geolocation data (SZA) were available. Uncertainties in the time series of PAR and GHI data were found to be equal to $\pm 6.05\%$ and $< \pm 7\%$ respectively. The generic performance of the models was then validated against satellite-corrected PAR and GHI calibrated data at 5 geographically distinct (and distant) stations across Greece. The major findings of this study are that:

- Quality assured calibrated data can be provided by the PAR channel of NILU-UV instruments and for GHI from CM pyranometers
- It is possible to construct accurate and generic noise models of PAR from GHI (and vice-versa) with quantifiable uncertainties based on the calibrated data obtained from RTM runs and satellite data
- Multilinear and nonlinear NN models in particular, allow aerosol and cloud effects expressed through the AOD and WV to be easily and implicitly incorporated
- Key to the performance of all linear and nonlinear NN models is M-SSA. For both PAR and GHI models, the reconstructed outputs based on time-delay embedding with 50 lag series correlate strongly with the target satellite-corrected data ($0.998 \leq r \leq 0.999$) and have very low reconstruction errors ($0.16\% \leq \text{MAPE} \leq 0.62\%$) despite the fact that there is an intrinsic error due to unaccounted-for variance in the truncated SVD. While the nonlinear NN model only marginally out-performs the linear LR and MLR models in this study, this small improvement of nonlinear models over linear models is attributed to the fact that the noise component represents only a small fraction of the time series' variance (< 4.1%) – i.e. signal to noise ratios of the order 23:1. Cloud effects appear to impact the “background” clear sky curve in this domain and are coincident with the noise component extracted by M-SSA.
- Provided that M-SSA is used to regularize the time series, NN models out-perform linear models of both the PAR and GHI with an average MAE with respect to target calibrated measurements $\approx 8.2 \pm 1.7 \text{ W/m}^2$ for PAR and $\approx 19.2 \pm 4.2 \text{ W/m}^2$ for GHI (well within the uncertainty of the measurements themselves), and an absolute mean bias $\approx 1.2 \text{ W/m}^2$ for PAR and $\approx 0.8 \text{ W/m}^2$ for GHI models. Both PAR and GHI models correlate strongly with calibrated measurements and achieved Pearson correlation coefficients $r > 0.978$ at all sites and coefficients of determination $R^2 > 0.957$.

In conclusion, the models of the PAR and GHI, once trained, produce fairly accurate outputs directly from matrices of input vectors. As such, computer processing is reduced to the problem of pre-processing the inputs with M-SSA which is not resource intensive. Furthermore, since PAR and GHI are central to contemporary studies in many biological and physical fields (e.g. renewable energy, agricultural productivity, carbon dynamics etc., the generic models developed here should be

able to assist studies in these areas by helping provide accurate long-term time series via time series gap-filling or direct and unsupervised modelling. The fairly high performance and generality of the models presented here means that they are to be recommended for widespread application. These technically sophisticated (albeit straight-forward to code) models can be used as reference predictions in other models like Numerical Weather Prediction (NWP) models or stochastic models, while long-term gap-filled datasets accompanied by specifiable uncertainties, can help create a benchmark against which model estimates can be validated.

Acknowledgements

M.M. Zempila acknowledges funding by the Greek State Scholarships Foundation (GRANT No: SR 22088/13) and Siemens AG under the “IKY-Siemens Post-doctoral Scholarship Programme 2013-2014”. The authors acknowledge the General Secretariat for Research and Technology, Greek Ministry of Education, Lifelong Learning and Religious Affairs for funding the project Hellenic Network of Solar Energy (www.helionet.gr, 09SYN-32-778) and the project “Greek National Network for UV Monitoring” funded by the General Secretariat for Research and Development (www.uvnet.gr, ΦΠ1 - 15100/17.11.2003). MeteoSwiss is acknowledged for providing the instruments for the establishment of the HNSE ground-based network. We also acknowledge OMI and MODIS mission scientists and associated personnel for the production of the satellite data in this study. M. Taylor would like to thank the members of LAP for their hospitality and technical support, Theodore Alexandrov for making available online his AutoSSA MATLAB code (<http://www.pdmi.ras.ru/~theo/autossa/>) and Nina Golyandina and co-workers for providing detailed open access resources on the Caterpillar-SSA method (<http://www.gistatgroup.com/cat/>).

References

- Aguiar L.J.G., GR. Fisher, RL. Malhado and RD. Aguiar: Modeling the photosynthetically active radiation in South West Amazonia under all sky conditions, *Theor. Appl. Climatol.*, 108, 631-640, 2012.
- Alados, I., I. Foyo-Moreno, and L. Alados-Arboledas: Photosynthetically Active Radiation: measurements and modeling. *Agric. For. Meteorol.* 78, 121-131, 1996.
- Alados-Arboledas, L., F.J. Olmo, I. Alados and M. Perez: Parametric models to estimate photosynthetically active radiation in Spain. *Agric. For. Meteorol.* 101, 187-201, 2000.
- Badescu, V.: *Modeling Solar Radiation at the Earth's Surface*. Springer, 2014.
- Bai, J. Observations and estimations of PAR and solar visible radiation in North China. *J. Atmos.*, 2012.
- Bais, A., Kazantzidis, A., Zerefos, C., Melas, D ;, Kosmidis, E., Kazadzis, S., Nikitidou, E., Giannaros, T.M., Zempila, M.M., Fragkos, K., Salamalikis, V. Hellenic network for Solar Energy. *Advances in Meteorology, Climatology and Atmospheric Physics*, Springer Atmospheric Sciences, Springer Berlin Heidelberg, http://dx.doi.org/10.1007/978-3-642-29172-2_55, 2013.
- Bais, A.F., Drosoglou Th., Meleti C. and Kouremeti, N.: Changes in Total Solar Irradiance Measured at Thessaloniki, Greece Since 1993 Associated with Changes in Aerosols. *Advances in Meteorology, Climatology and Atmospheric Physics*, http://dx.doi.org/10.1007/978-3-642-29172-2_125, 2013.
- Bais, A.F., S. Kazadzis, D. Balis, C.S. Zerefos, and M. Blumthaler Correcting Global Solar Ultraviolet Spectra Recorded by a Brewer Spectroradiometer for its Angular Response Error. *Appl. Opt.* Vol. 37, Issue 27, pages 6339-6344, doi: 10.1364/AO.37.006339, 1998.
- Beale, M.H., Hagan, M.T., Demuth, H.: *Neural Network Toolbox: User's Guide*, The MathWorks, Inc., Natick, MA (USA), 2015.
- Bishop, C. M.: Training with noise is equivalent to Tikhonov regularization, *Neural Comput.*, 7, 108–116, doi:10.1162/neco.1995.7.1.108, 1995.
- Bosch, J.L., G. Lopez and F.J. Batles: Global and direct photosynthetically active radiation parameterizations for clear-sky conditions. *Agric. Meteorol.* doi: 10.1016/j.agrformet.2008.07.011, 2008.

- Briceño, H., Rocco, C. M., & Zio, E.: Singular spectrum analysis for forecasting of electric load demand. In Prognostics and System Health Management Conference, Vol 33, 919-924, 2013.
- Broomhead, D.S., and G.P. King: "Extracting qualitative dynamics from experimental data", Physica D, 20, 217–236, 1986.
- Burnett, D., Barbour, E., Harrison, G.P.: The UK solar energy resource and the impact of climate change, Renewable Energy, Volume 71, Pages 333-343, ISSN 0960-1481, <http://dx.doi.org/10.1016/j.renene.2014.05.034>, 2014.
- Cancillo, M.L., A. Serrano, M. Anton, J.A. Garcia and J.M. Vilaplana: Improving the solar zenith angle dependence of broadband UV radiometers calibration. Atmos. Chem. Phys. Discuss., 7, 17873–17892, 2007.
- Cho, H.K., M.J. Jeong, J. Kim and Y.J. Kim: Dependence of diffuse photosynthetically active solar irradiance on total optical depth. J. Geophys. Res. 108, D9, 4267, 4-1-4-10, 2003.
- Costa MH, Yanagi SNM, Souza P, Ribeiro A, Rocha EJP: Climate change in Amazonia caused by soybean cropland expansion, as compared to caused by pastureland expansion. Geophys Res Lett 34(7), 2007.
- Cybenko, G.: Approximation by super-positions of a sigmoidal function. Math. Contro. Signals 2: 303–314, 2989.
- Dahlback A., H.A. Eide, B.A. K. Hoiskar, R.O. Olsen, F.J. Schmidlin, S-C. Tsay, K. Stamnes: Comparison of data for ozone amounts and ultraviolet doses obtained from simultaneous measurements with various standard ultraviolet instruments, Opt. Eng. 44(4), 041010, doi:10.11117/1.1885473, 2005.
- Dahlback, A.: Measurements of biologically effective UV doses, total ozone abundances, and cloud effects with multichannel, moderate bandwidth filter instruments. Appl. Opt., 35, 6514-6521, 1996.
- Feister U., R. Grewe and K. Gericke: A method for correction of cosine errors in measurements of spectral UV irradiance. Solar Energy, 60, 313-332, doi:10.1016/S0038-092X(97)00030-3, 1997.
- Fernández-Peruchena, C.M. and M. Gastón: A simple and efficient procedure for increasing the temporal resolution of global horizontal solar irradiance series. Renewable Energy, Volume 86, 375-383,0960-1481, doi: 10.1016/j.renene.2015.08.004, 2016.
- Frouin R. and R.T. Pinker: Estimating Photosynthetically Active Radiation (PAR) at the Earth's atmosphere from satellite observations. Remote Sens. Environ. 51, 98-107, 1995.

Ge, S., Smith, R.G., Jacovides, C.P., Kramer, M.G. and Carruthers, R.I.: Dynamics of photosynthetic photon flux density (PPFD) and estimates in coastal northern California. *Theoretical and applied climatology*, 105(1-2), pp.107-118, 2011.

Ghil, M., Allen, M. R., Dettinger, M. D., Ide, K., Kondrashov, D., Mann, M. E. and Yiou, P.: Advanced spectral methods for climatic time series. *Reviews of geophysics*, 40(1), 3-1, 2002.

Golyandina, N., & Korobeynikov, A.: Basic singular spectrum analysis and forecasting with R. *Computational Statistics & Data Analysis*, 71, 934-954, 2014.

Golyandina, N., Nekrutkin, V. and Zhigljavsky, A.: Analysis of time series structure: SSA and related techniques, Chapman and Hall/CRC. ISBN 1-58488-194-1, 2001.

Gonzalez JA. and Calbo J.: Modeled and measured ratio of PAR to global radiation under cloudless skies. *Agric. For. Meteorol.*, 110, 319-325, 2002.

Gueymard, C. A. and Myers, D. R.: Validation and ranking methodologies for solar radiation models. In *Modeling Solar Radiation at the Earth's Surface* (pp. 479-510). Springer Berlin Heidelberg, 2008.

Gueymard, C.: A two-band model for the calculation of clear sky solar irradiance, illuminance, and photosynthetically active radiation-Validation with a benchmark dataset. *Sol. Energy* 43, 253-265, 1989.

Gueymard, C.: REST2: high-performance solar radiation model for cloudless-sky irradiance illuminance, and photosynthetically active radiation-Validation with a benchmark dataset. *Sol. Energy* 82, 272-285, 2008.

Hagan, M. T. and Menhaj, M. B.: Training feedforward networks with the Marquardt algorithm, *IEEE T. Neural Netw.*, 5, 989–993, doi:10.1109/72.329697, 1994.

Hanan, N.P. and A. Begue: A method to estimate instantaneous and daily intercepted photosynthetically active radiation using a hemispherical sensor. *Agric. For. Meteorol.* 74, 155-168, 1995.

Hassani, H., and Mahmoudvand, R.: Multivariate Singular Spectrum Analysis: A General View and New Vector Forecasting Approach, *International Journal of Energy and Statistics* 1(1), 55-83, 2013.

Hassani, H., Heravi, S., & Zhigljavsky, A.: Forecasting European industrial production with singular spectrum analysis. *International journal of forecasting*, 25(1), 103-118, 2009.

Høiskar, B.A. K., R. Haugen, T. Danielsen, A. Kylling, K. Edvardsen, A. Dahlback, B. Johnsen, M. Blumthaler and J. Schreder: Multichannel Moderate-Bandwidth Filter Instrument for Measurement of the Ozone-Column Amount, Cloud Transmittance, and Ultraviolet Dose Rates. *Applied Optics*, Vol. 42, Issue 18, pp. 3472-3479, doi: 10.1364/AO.42.003472, 2003.

Holmström, L. and Koistinen, P.: Using additive noise in backpropagation training. *IEEE T. Neural Networks*, 3, 24–38, doi:10.1109/72.105415, 1992.

Hornik, K., Stinchcombe, M. and White, H.: Multilayer Feedforward Networks Are Universal Approximators. *Neural Networks*, 2, 359-366, 1989.

IEA: Electricity Information 2012. Published by the International Energy Agency, 883, 2012.

IEA: World Energy Outlook 2014, Executive Summary, Published by the International Energy Agency, 2014.

Janjai S, Wattan R. Development of a model for the estimation of photosynthetically active radiation from geostationary satellite data in a tropical environment. *Remote Sens Environ* 2011; 115(7): 1680 -1693, 2011.

Kaiser, F.G.: System Analysis by Digital Computer. Wiley, New York, 1974.

Kaiser, H.F.: The varimax criterion for analytic rotation in factor analysis. *Psychometrika* 23, 187-200, 1958.

Kandasamy, S., Baret, F., Verger, A., Neveux, P., & Weiss, M.: A comparison of methods for smoothing and gap filling time series of remote sensing observations—application to MODIS LAI products. *Biogeosciences*, 10(6), 4055-4071, 2013.

Kashyap, Y., A. Bansal and A.K. Sao: Solar radiation forecasting with multiple parameters neural networks. *Renewable and Sustainable Energy Reviews*, Volume 49, 825-835, doi: 10.1016/j.rser.2015.04.077, 2015.

Kirk, J. T. O.: Spectral distribution of photosynthetically active radiation in some south-eastern Australian waters. *Marine and Freshwater Research*, 30(1), 81-91, 1979.

Lakkala K., Jaros A., Aurela M., Tuovinen J.-P., Kivi R., Suokanerva H., Karhu J.M. & Laurila T.: Radiation measurements at the Pallas-Sodankylä Global Atmosphere Watch station — diurnal and seasonal cycles of ultraviolet, global and photosynthetically-active radiation. *Boreal Env. Res.* 21: 00–00, 2016.

Landelius, T. and W. Josefsson: Methods for cosine correction of broadband UV data and their effect on the relation between UV irradiance and cloudiness. *JGR* 105, doi: 10.1029/1999JD900982, 2000.

- Landsberg, J. J., Waring, R. H.: A generalised model of forest productivity using simplified concepts of radiation-use efficiency, carbon balance and partitioning. *Forest Ecology and Management*, 95(3), 209-228, 1997.
- Levenberg, K.: A Method for the Solution of Certain Non-Linear Problems in Least Squares. *Quarterly of Applied Mathematics* 2: 164–168, 1944.
- Long, M. H., Rheuban, J. E., Berg, P., & Zieman, J. C.: A comparison and correction of light intensity loggers to photosynthetically active radiation sensors. *Limnology and Oceanography: Methods*, 10(6), 416-424, 2012.
- G. López G., M.A. Rubio, M. Martínez, F.J. Batllés: Estimation of hourly global photosynthetically active radiation using artificial neural network models. *Agricultural and Forest Meteorology*, 107, 279–291, 2001.
- Ma S., G. Churkina and Trusilova K.: Investigating the impact of climate change on crop phenological events in Europe with a phenology model. *Int. J. Biometeorol.*, 56, 749-763, 2012.
- Marquardt, D.: An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *SIAM Journal on Applied Mathematics* 11 (2): 431–441, 1963.
- Mayer B. and Kylling A.: Technical note: The libRadtran software package for radiative transfer calculations - description and examples of use, *Atmos. Chem. Phys.*, 5, 1855-1877, doi: 10.5194/acp-5-1855-2005, 2005.
- Mayer H., T. Holst and D. Schindler: Microclimate within beech stands-part1: photosynthetically active radiation. *Forstw Cbl* 121, 301-321, 2002.
- McCandless, T.C., S.E. Haupt and G.S. Young: A model free approach to forecasting solar irradiance variability. *Solar Energy*, Volume 120, 514-524, doi: 10.1016/j.solener.2015.07.020, 20015.
- McCree K: Test of current definitions of photosynthetically active radiation against leaf photosynthesis data. *Agric Meteorol* 10:443–453, 1972.
- Møller, M. F.: A scaled conjugate gradient algorithm for fast supervised learning, *Neural Networks*, 6, 525–533, doi:10.1016/S0893-6080(05)80056-5, 1993.
- Pedro,H. and C. Coimbra: Nearest-neighbor methodology for prediction of intra-hour global horizontal and direct normal irradiances. *Renewable Energy*, Volume 80, 770-782, doi: 10.1016/j.renene.2015.02.061, 2015.

Pei FS., X. Li, XP. Liu and CH. Lao: Assessing the impacts of droughts on net primary productivity in China. *J. Environ. Manage.*, 114, 362-371, 2013.

Perry, S. G., Fraser, A. B., Thomson, D. W., & Norman, J. M.: Indirect sensing of plant canopy structure with simple radiation measurements. *Agricultural and Forest meteorology*, 42(2), 255-278, 1988.

Pierro, M., F. Bucci, C. Cornaro, E. Maggioni, A. Perotto, M. Pravettoni and F. Spada: Model output statistics cascade to improve day ahead solar irradiance forecast. *Solar Energy*, Volume 117, 99-113, doi: 10.1016/j.solener.2015.04.033, 2015.

Ross, J. and M. Sulev: Sources of errors in measurements of PAR. *Agric. For. Meteorol.* 100, 103-125, 2000.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J.: Learning representations by back-propagating errors, *Nature*, 323, 533–536, doi:10.1038/323533a0, 1986.

Salas, V., Olias, E.: Overview of the photovoltaic technology status and perspective in Spain. *Renew. Sustain. Energy Rev.* 13 (5), 1049–1057, 2009.

Sarralde, J.J., Quinn, D.J., Wiesmann, D. and Steemers K.: Solar energy and urban morphology: Scenarios for increasing the renewable energy potential of neighbourhoods in London, *Renewable Energy*, Volume 73, Pages 10-17, ISSN 0960-1481, <http://dx.doi.org/10.1016/j.renene.2014.06.028>, 2015.

Seckmeyer, G., G. Bernhard: Cosine error correction of spectral UV-irradiances. *Proc. SPIE 2049, Atmospheric Radiation*, 140, doi:10.1117/12.163505, 1983.

Sudharsan D., J. Adinarayana, DR. Reddy, G. Screenivas, S. Ninomiya and SN. Merchant: Evaluation of weather-based rice yield models in India. *Int. J. Biometeorol.*, 57, 107-123, 2013.

Sundaram, S. and Jakka Sarat Chandra Babu: Theoretical prediction and validation of global horizontal solar irradiance for a tropical climate in India, *Frontiers in Energy*, Volume 9, Issue 3, 311-321, doi: 10.1007/s11708-015-0369-3, 2015.

Takens, F.: Detecting strange attractors in turbulence. In: *Lect. Not. in Math.*, vol. 898, *Dynamical Systems and Turbulence*, edited by D. A. Rand and L.-S. Young, pp. 366 –381, Springer-Verlag, New York., 1981.

Taylor, M., Kazadzis, S., Tsekeri, A., Gkikas, A., Amiridis, V.: Satellite retrieval of aerosol microphysical and optical parameters using neural networks: a new methodology applied to the Sahara desert dust peak. *Atmos. Meas. Tech.* 7, 3151–3175. doi:10.5194/amt-7-3151-2014, 2014.

- Taylor, M., Kosmopoulos, P.G., Kazadzis, S., Keramitsoglou, I., Kiranoudis, C.T.: Neural network radiative transfer solvers for the generation of high resolution solar irradiance spectra parameterized by cloud and aerosol parameters. *J. Quant. Spectr. Rad. Trans* 168(C), 176-192. doi:10.1016/j.jqsrt.2015.08.018, 2016.
- Tsubo, M., & Walker, S.: A model of radiation interception and use by a maize-bean intercrop canopy. *Agricultural and Forest Meteorology*, 110(3), 203-215, 2002.
- Udo S.O and T.O. Aro: Global PAR related to global solar radiation for central Nigeria. *Agric. For. Meteorol.* 97, 21-31, 1999.
- Utrillas, M.P., M.J. Marin, A.R. Esteve, F. Tena, J. Canada, V. Estelles and J.A. Martinez-Lozano: Diffuse UV erythemal radiation experimental values. *JGR*, VOL. 112, D24207, doi:10.1029/2007JD008846, 2007.
- Vautard, R., and M. Ghil.: Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series, *Phys. D*, 35, 395–424, 1989.
- Wang, B. and GY Shi: Long-term trends of atmospheric absorbing and scattering optical depths over China region estimated from the routine observation data of surface solar irradiances. *J. Geophys. Res.*, 115, D00K28, doi: 10.1029/2009JD013239, 2010.
- Wang, LC., W. Gong, C. Li, AW. Lin, B. Hu and YY MA: Measurement and estimation of photosynthetically active radiation from 1961 to 2011 in Central China. *Appl. Energ.*, 111, 1010-1017, 2013.
- Wang, L., Gong, W., Ma, Y., Hu, B., & Zhang, M.: Photosynthetically active radiation and its relationship with global solar radiation in Central China. *International journal of biometeorology*, 58(6), 1265-1277, 2015a.
- Wang, L., Gong, W., Hu, B., Lin, A., Li, H. and Zou, L.: Modeling and analysis of the spatiotemporal variations of photosynthetically active radiation in China during 1961–2012. *Renew. Sustain. Energy Rev.* 49:1019-1032, 2015b.
- Willmott, C. J., Ackleson, S. G., Davis, R. E., Feddema, J. J., Klink, K. M., Legates, D. R. and Rowe, C. M.: Statistics for the evaluation and comparison of models. *Journal of Geophysical Research: Oceans* (1978–2012), 90(C5), 8995-9005, 1985.
- Wilson, D. R. and Martinez, T. R.: The inefficiency of batch training for large training sets, in: *Proc. IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN 000)*, Vol. 2, 113–117, doi:10.1109/IJCNN.2000.857883, 2002.

- Xu, J., Bergin, M. H., Greenwald, R., & Russell, P. B.: Direct aerosol radiative forcing in the Yangtze delta region of China: Observation and model estimation. *Journal of Geophysical Research: Atmospheres* (1984–2012), 108(D2), 2003.
- Young, S., David, C. T., Gibson, G.: Light measurement for entomology in the field and laboratory. *Physiological Entomology* (1987) 12,373-379.
- Zagouras, A., A. Kazantzidis, A., Nikitidou, E., Argiriou, A.A.: Determination of measuring sites of solar irradiance, based on cluster analysis of satellite-derived cloud estimations. *Solar Energy*, 97, 1-11, 2013.
- Zagouras, A., Hugo T.C. Pedro, Carlos F.M. Coimbra: Clustering the solar resource for grid management in island mode, *Solar Energy* 110 (2014) 507–518, 2014.
- Zell, E., Ss Gasim, S. Wilcox, S. Katamoura, T. Stoffel, H. Shibli, J. Engel-Cox and M. Al Subie: Assessment of solar radiation resources in Saudi Arabia. *Solar Energy*, Volume 119, 422-438, doi: 10.1016/j.solener.2015.06.031, 2015.
- Zempila, M.M., T.M. Giannaros, A. Bais, D. Melas and A. Kazantzidis: Evaluation of WRF shortwave radiation parameterizations in predicting Global Horizontal Irradiance in Greece. *Renewable Energy*, Volume 86, 831-840,doi: 10.1016/j.renene.2015.08.057, 2016a.
- Zempila, M.M., Taylor, M., Fountoulakis, I., Bais, A., Kazadzis, S., Frakos, K.: Introducing a cloud screening detector using global horizontal irradiances in UV and PAR in Thessaloniki, Greece. 13th International Conference on Meteorology, Climatology & Atmospheric Physics (COMECP), 19-21 September 2016, Thessaloniki, Greece, 2016b.
- Zhang X, Y. Zhang and Y. Zhaou: Measuring and modeling photosynthetically active radiation in Tibetan Plateau during April-October. *Agric. For. Meteorol.*, 102, 207-212, 2000.

Appendix A

The starting point of PCA-FA is to perform two initial factor analysis tests (Kaiser, 1974) on the full matrix of input variables: 1) Barlett's Test for sphericity, and 2) categorization of the data using the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy. In order to apply these tests, we assembled the co-located matrix of input variables (see Eq. 2 and Eq. 3) using available data at the Thessaloniki site which is to be used for construction of the GHI and PAR models. The number of complete data records is N=47,088. For both models, Barlett's sphericity test was found to show a very high degree of relationship between the 4 variables with χ^2 values in excess of 10^5 and p-values << 0.005, suggesting that the data is suitable for factor analysis at the 95% confidence level. However, the KMO measure of sampling adequacy (which is in the range 0 to 1) was found to be very low with values less than 0.5 (0.48 for the GHI model inputs and 0.45 for the PAR model inputs) - suggesting that the degree of common variance is “unacceptable” for factorization according to the categorization of Kaiser (1974). In such a case of an initial contradiction, further initial information can be gained by looking at the correlation between variables for both sets of model inputs. Analysis of the covariance matrix of inputs (centered and normalized values, i.e. z-scores) indicated the presence of a strong anti-correlation between SZA and PAR ($R = -0.983$), and SZA and GHI ($R = -0.981$), a very weak anti-correlation between SZA and WV ($R = -0.280$), and a mildly positive correlation between WV and AOD ($R = 0.491$) for both input matrices. In order to assess whether or not the existence of these correlations is sufficiently strong to warrant or require dimensional reduction, we performed standard PCA and also PCA with VARIMAX rotation (Kaiser, 1958) to the input matrix. The results are presented in Table A1.

Table A1 The results of application of standard PCA (unrotated eigenvectors) and PCA with VARIMAX rotation to the matrix of input data for the GHI and PAR models.

GHI eigenvalues (unrotated)	Variance (%)	Cumulative (%)	GHI eigenvalues (VARIMAX rotated)	Variance (%)	Cumulative (%)
2.13	53.26	53.26	1.98	49.50	49.50
1.38	34.46	87.72	1.01	25.21	74.71
0.48	11.98	99.70	1.00	24.99	99.69
0.01	0.30	100.00	0.01	0.31	100.00
PAR eigenvalues (unrotated)	Variance (%)	Cumulative (%)	PAR eigenvalues (VARIMAX rotated)	Variance (%)	Cumulative (%)
2.11	52.82	52.82	1.98	49.45	49.45

1.39	34.85	87.67	1.01	25.21	74.66
0.48	12.04	99.71	1.00	25.05	99.71
0.01	0.29	100.00	0.01	0.29	100.00

A typical guideline, provided by VARIMAX-rotated PCA, is that the number of principal components (PCs) is indicated by the number of eigenvalues greater than unity (i.e. 3 in our case). This finding is echoed also by the factor loading (or “weight matrix”) of the variance explained in each variable by each principal component (PC) in both input matrices – which also suggests that 3 of the 4 PCs contribute substantially to at least one of the input variables as shown in Fig. A1a) and Fig. A1b). For example, while the contribution to the variance of the GHI and PAR from PC4 seems insignificant (as it only explains 0.6% of the variance), it is clear from Fig. A1c) and Fig. A1d) that all 4 PCs are needed to reconstruct this variable’s variance fully.

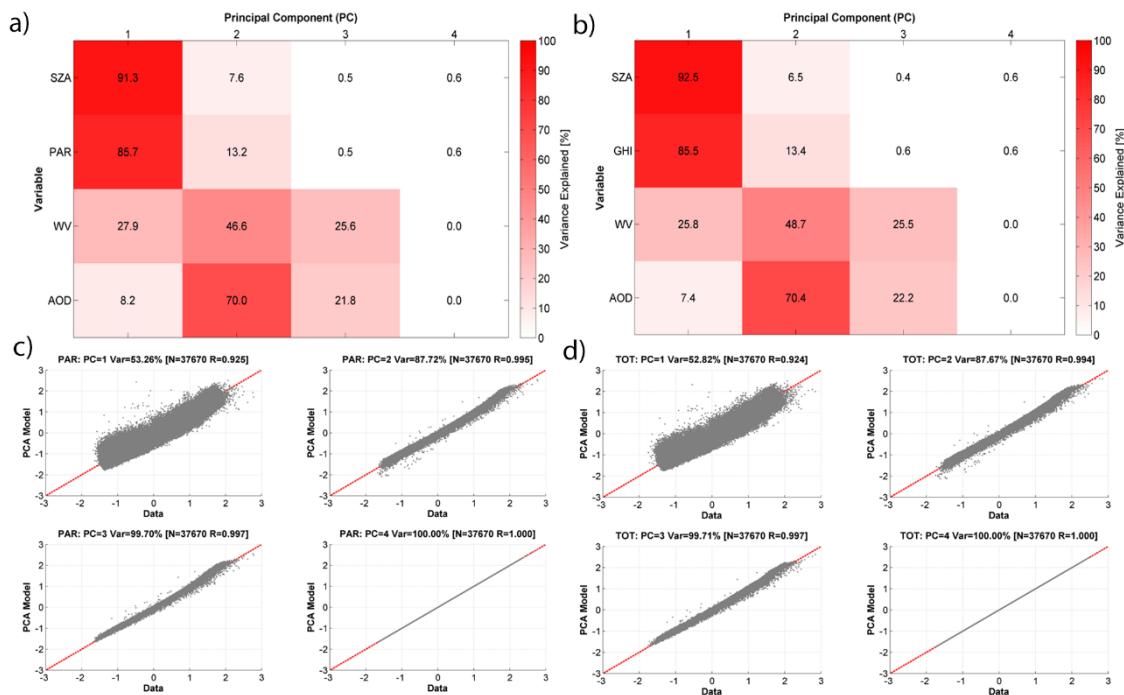


Figure A1 The factor loading as a percentage of the variance of each input variable accounted for by each PC for a) GHI model inputs and b) PAR model inputs. Plots c) and d) show the regression of the PCA model of the GHI and PAR irradiance input variables respectively on the raw time series data.

Appendix B

Analysis of the performance of multivariate models like the PAR and GHI models presented below, inevitably involves a choice of the statistics used to assess the goodness of fit of model simulations or predictions (P_j) against measurements or observations (O_j). As a result of the work of Willmott et al. (1985) which outlined a set of difference measures that can be used to evaluate the operational performance of a wide spectrum of geophysical models (regardless of whether the model predictions are manifested as scalars, directions, or vectors), difference or error measures have gradually replaced correlation- and skill-based indices which are not consistently linked to model accuracy. In particular, Willmott et al. (1985) demonstrated that both model accuracy and precision can be meaningfully estimated by analyzing and reporting the following narrow but extensive set of data versus model statistics. Firstly, at the level of exploratory statistics, the average values of predictions and observations given by the means μ_p and μ_o and their variability by the standard deviations s_p and s_o should be reported. The reason for this is that in scatterplots containing a large number of model outputs and measurement data, a cloud of points is often observed. The means μ_p and μ_o tell us where the centre of such a cloud is and the deviations s_p and s_o measure how dispersed it is around this centre. Secondly, for differences $D_j = P_j - O_j$ the mean bias (b) calculated via $b = \sum_{j=1}^N D_j / N$ should be used to measure the direction of the average error (i.e. whether or not there is under- or over-estimation), while the mean absolute error (MAE) calculated from:

$$E^{1/\gamma} = \left[\frac{\sum_{j=1}^N |D_j|^\gamma}{N} \right]^{1/\gamma} \quad (B1)$$

with $\gamma = 1$ should be used to measure its size. Then, setting $\gamma = 2$ in Eq. B1, the root mean-squared error (RMSE) provides a complimentary measure (to the MAE) of the size of the average error that is more sensitive (being calculated from squared differences). In addition, Willmott et al. (1985) demonstrated that both model accuracy and precision can be meaningfully estimated by analyzing and reporting also the systematic average error:

$$RMSE_s = \sqrt{\frac{\sum_{j=1}^N |\hat{P}_j - O_j|^2}{N}} \quad (B2)$$

and the *unsystematic* average error:

$$RMSE_u = \sqrt{\frac{\sum_{j=1}^N |\hat{P}_j - P_j|^2}{N}} \quad (B3)$$

calculated by decomposing the RMSE using \widehat{P} , an OLS estimate of P obtained by regressing P on O. The RMSE_s, (like the MAE and the RMSE) is also a (more sensitive) measure of the size of the linear bias between model predictions and observations, while RMSE_u is a measure of model precision (Willmott et al., 1985). Note that, being dependent on the differences $D_j = P_j - O_j$ the quantities b, MAE, RMSE, RMSE_s and RMSE_u all have the same units as the data. Willmott et al. (1985) also describe two unit-independent indices of agreement d_1 and d_2 as measures of model performance in the interval (0,1) which should be reported and which depend on linear differences (d_1) and squared differences (d_2) given by:

$$d_\gamma = 1 - \frac{\sum_{j=1}^N |D_j|^\gamma}{\sum_{j=1}^N (|P_j - \mu_P| + |O_j - \mu_O|)^\gamma} \quad (\text{B4})$$

such that d_1 corresponds to $\gamma = 1$ and d_2 corresponds to $\gamma = 2$. Finally, as a measure of the amount of variance captured by the model, Willmott et al. (1985) recommend also calculation of the coefficient of determination R^2 . In this work therefore, we adopt this narrow but extensive set of statistics to evaluate the performance of the generic NN models. Another important point to note here is that, while we report measurement errors in the form of the MAPE, when calculating the statistics above in the context of data versus NN model comparisons, we follow the recommendation of Willmott et al. (1985) that observations are assumed to be effectively “error-free” so that the statistical framework described above is valid and can be applied. When presenting data versus model statistics we will distinguish between i) exploratory statistics (means and standard deviations of the measurements and model outputs), ii) linear (1st moment) statistics (b, MAE, r and d_1) and iii) quadratic (2nd moment) statistics (RMSE, RMSE_s, RMSE_u, R^2 and d_2) which are more sensitive measures of difference.

Highlights

- Generic linear regression and nonlinear neural network models of PAR and GHI are constructed
- The LibRadtran UVspec model is used to calibrate PAR data and convert counts to irradiance
- Multi-channel singular spectrum analysis is central to regularization of PAR and GHI time series
- PAR and GHI models use solar zenith angle, aerosol optical depth and water vapor as inputs
- Nonlinear neural network models are found to perform slightly better than linear models