UiT
THE ARCTIC
UNIVERSITY
OF NORWAY

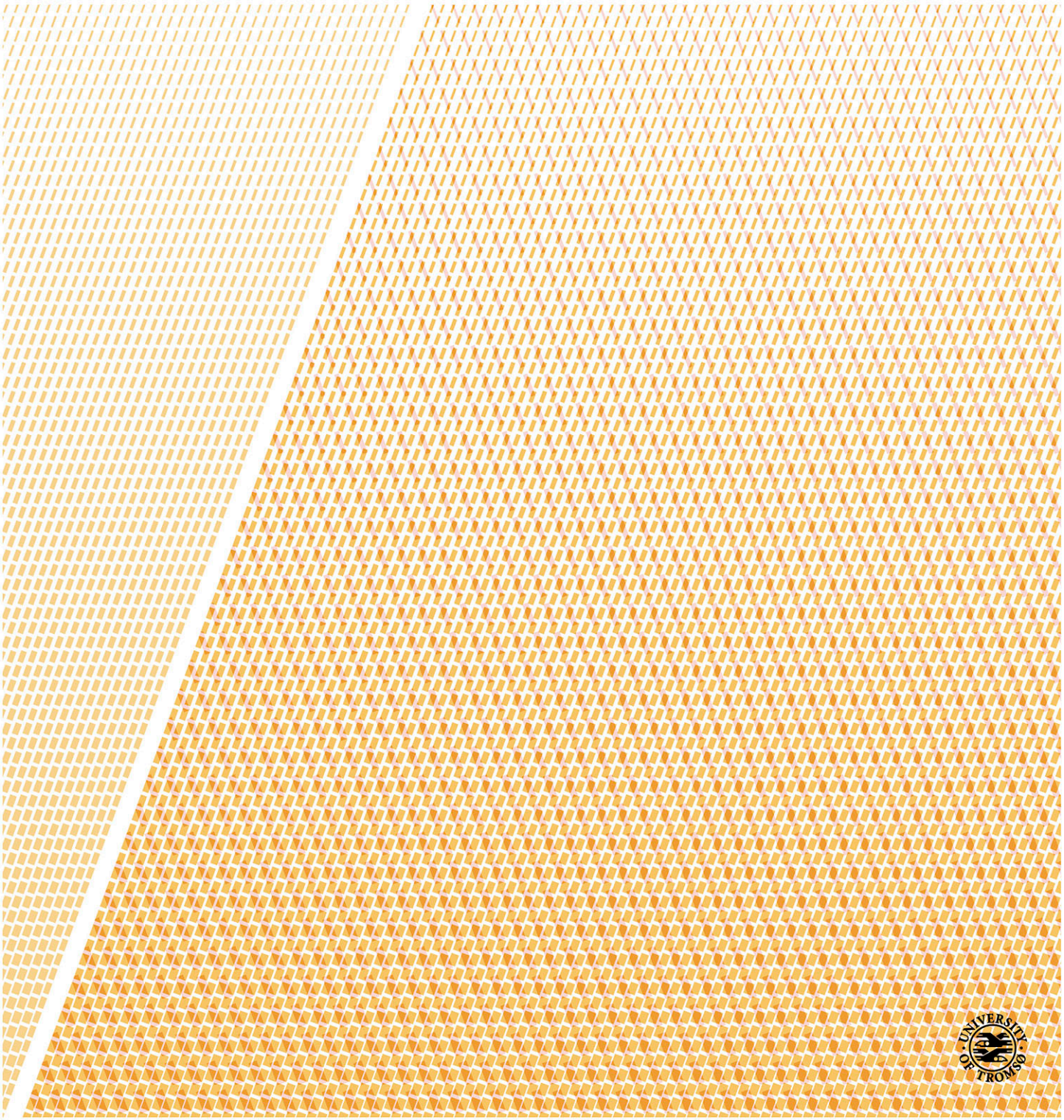
Faculty of Science and Technology
Department of Physics and Technology

Solar resource assessment at high latitude regions

—

Bilal Babar

A dissertation for the degree of Philosophiae Doctor – January 2019



‘You're the only reason I am… you are all my reasons.’

-John Forbes Nash, Jr. (June 13, 1928 — May 23, 2015)

The efforts required in creating any document far exceed the capabilities of any one person. This thesis has been no exception.

Abstract

There has been a growing interest in accurately estimating surface solar radiation at high latitude locations. From a Scandinavian perspective, the installed solar photovoltaic share is increasing, primarily because of the declining cost of these systems, the introduction of various economic incentives and societal push to generate one's own clean power. In the coming years, it is anticipated that the share of photovoltaics in the energy mix of Scandinavia will increase substantially.

One of the main deterrent in an accurate estimation of surface solar radiation is the limited coverage of geostationary satellites. These satellites, which are widely used globally to estimate solar radiation, do not provide coverage above 65ºN. Alternatively, polar orbiting satellites can be used to estimate surface solar radiation but a low sensing frequency and difficulties in differentiating clouds from snow-covered surfaces result in a large number of missing values in the data. Moreover, reanalyses also provide surface solar radiation estimates and in recent years, it is seen that the accuracy of reanalyses with respect surface solar radiation is getting better.

This thesis starts with providing an evaluation and comparative analyses of different solar radiation datasets for high latitude locations. First, an empirical model based on intra-day temperature differences and relative humidity is proposed. This model can be used at meteorological stations that do not have dedicated equipment to estimate surface solar radiation. Then, a comparative analysis is performed for Norwegian locations in which four different models were evaluated. It was found that satellite databases are more accurate than reanalyses and empirical models. However, satellite databases were found to underestimate solar radiation while reanalyses were found to overestimate it. After this, a study was performed to evaluate the CLARA-A1 and CLARA-A2 polar orbiting satellite based datasets. It was found that the CLARA-A2 dataset has less number of missing values but mostly the increase in data is at snow-covered surfaces. The data in CLARA-A2 has higher accuracy than CLARA-A1, but at these new data points which were previously not available in CLARA-A1, the errors are very large.

Finally, a novel regression-based solar radiation dataset is proposed here that uses one polar orbiting satellite dataset (CLARA-A2), one global reanalysis (ERA5), and auxiliary data based on Sun-Earth geometric relationships. The proposed dataset has better accuracy and precision than CLARA-A2 and ERA5 datasets.

List of papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I. Babar B, Boström T. Estimating solar irradiation in the Arctic. Renewable Energy and Environmental Sustainability. 2016;1:34.

II. Babar B, Graversen R, Boström T. Evaluating CM-SAF solar radiation CLARA-A1 and CLARA-A2 datasets in Scandinavia. Solar Energy. 2018 Aug 31;170:76-85.

III. Babar B, Graversen R, Boström T. Solar radiation estimation at high latitudes: Assessment of the CMSAF databases, ASR and ERA5 (accepted for publication in Solar Energy with minor revisions)

IV. Babar B, Luppino L T, Boström T, Anfinsen S N. Random forest regression for improved mapping of solar power resources at high latitudes (manuscript)

Publication not included in the thesis

V. Solbakken K, Babar B, Boström T. Correlation of wind and solar power in high-latitude arctic areas in Northern Norway and Svalbard. Renewable Energy and Environmental Sustainability. 2016;1:42.

Notes on my contribution

I contributed with the following in the appended papers:

Paper I: I did the modelling, performed the error analysis, and wrote most of the paper.

Paper II: I performed the analysis, produced the plots, and wrote most of the paper.

Paper III: I performed the analysis, produced the plots, and wrote most of the paper.

Paper IV: I contributed in the modelling, evaluated the model results, and contributed in writing the manuscript.

Table of Contents

Abbreviations, Nomenclature and Subscripts

Abbreviation	Description
3DVAR	Three-dimensional Variational Data Assimilation
4DVAR	Four-dimensional Variational Data Assimilation
AM	Air Mass
ASR	Arctic System Reanalysis
AU	Earth-Sun distance in Astronomical units
AVHRR	Advanced Very High Resolution Radiometer
BRDF	Bidirectional Reflectance Distribution Function
BSRN	Baseline Surface Radiation Network
CAL	Effective Cloud Albedo
CDR	Climate Data Record
CERES	Cloud and Earth's Radiant Energy System
CLARA	Cloud, Albedo, Radiation Dataset
CM-SAF	Satellite Application Facility on Climate Monitoring
DNI	Direct Normal Irradiance
ECMWF	European Center for Medium Range Weather Forecast
ERA5	ECMWF Reanalysis 5
GADS	Global Aerosol Data Set
GEWEX	Global Energy and Water Cycle Experiment
GHG	Greenhouse Gases
GHI	Global Horizontal Irradiance
GHI	Global Horizontal Irradiance
IDW	Inverse Distance Weighting
IFS	Integrated Forecast System
IPCC	International Panel on Climate Change
ISCCP	International Satellite Cloud Climatology Project
libRadtran	A radiative transfer model
MABD	Mean Absolute Bias Deviation
MACC	Monitoring Atmospheric Composition and climate project
MABD	Mean Absolute Bias Deviation
MAGIC	Mesoscale Atmospheric Global Irradiance Code

Abbreviation	Description
MBD	Mean Bias Deviation
MetOp	Meteorological Operational Satellite Programme
MSF	Meteosat First Generation
MSG	Meteosat Second Generation
Mtoe	Million tons of oil equivalent
MVIRI	Meteosat Visible Infra-Red Imager
NIBIO	Norwegian Institute of Bioeconomy Research
NOAA	National Oceanic and Atmospheric Administration
OPAC	Optical Properties of Aerosol and Cloud
PV	Photovoltaic
QC	Quality Control
RFR	Random Forest Regression
RMSD	Root Mean Square Deviation
RTM	Radiative Transfer Model
SAFNWC	Nowcasting SAF
SARAH	Surface Solar Radiation Dataset- Heliosat
SARB	Surface and Atmospheric Radiation Budget
SDI	Surface Direct Irradiance
SDU	Sunshine Duration
SEVIRI	Spinning Enhanced Visible and Infrared Imager
SIS	Surface Incoming Shortwave Radiation
SMHI	Swedish Meteorological and Hydrological Institute
So	Solar constant at mean Earth-Sun distance
SoDa	Solar Radiation Data
SRI	Spectrally Resolved irradiance
SZA	Solar Zenith Angle
T	Transmittance
TFC	Total Final Consumption
TPES	Total Primary Energy Supply
WMO	World Meteorological Organization
WRF	Weather Research and Forecast
WRMC	World Radiation Meteorological Center

1. Introduction

As the human population is increasing, so is the global energy requirement. The increase in the energy requirement has exerted an escalating pressure on the climate in the form of emitted greenhouse gases leading to global warming. In the past 200 years, the production of heat and electric energy has been mainly from fossil-based systems. Due to the increasing population and economic development, the energy consumption is increasing even though the amount of energy required to produce one unit of income has decreased because of the advancements in technology and innovation. To mitigate the effects of climate change, nowadays there is a global drive to move towards cleaner and safer renewable energy systems. In this regard, the solar photovoltaic (PV) systems that generate electric energy based on irradiance from the Sun are increasing rapidly as well. In 2017, solar PV installations generated over 460 TWh of energy, which represents around 2% of global power output. There has been a growing interest in solar PV in the Nordic regions, but due to high latitude and frequent snow covers, the estimation of surface solar radiation from remote sensing techniques is not straightforward in these regions. The motivation behind this thesis lies in assessing the existing methods to estimate surface solar radiation in high latitudes and to provide improvement strategies for a better estimation of solar radiation in these regions. The lack of published research in this area represents a significant knowledge gap; the outcome of this thesis and appended papers is intended to give the scientists and policy makers a better understanding of surface solar radiation at high latitudes. This thesis starts with the assessment of available solar radiation sources like satellite and reanalysis, and concludes by proposing a regression method that significantly improves the estimated surface solar radiation.

1.1 Aim of the thesis

The central aim of this thesis is to analyze existing models that estimate surface solar radiation and to propose methods that can improve the current models for high-latitude locations. Estimating surface solar radiation from satellites is a well-developed and widely used method. On the other hand, reanalyses also provide surface solar radiation in addition to a number of other meteorological variables. A Reanalysis is based on data assimilation of observations and model-based forecasts, to estimate weather conditions. Solar radiation estimates from reanalyses are not as accurate as those obtained from satellite methods, but some recent studies have shown that the solar radiation estimates from reanalysis are improving and these can be

used to fill the missing values in satellite databases. This thesis has the following specific aims, which are addressed in the appended papers:

- *Developing a mathematical model to estimate surface solar radiation by using meteorological variables (Paper I).*
- *Analyzing the improvements in the recent polar-orbiting-satellite based datasets (Paper II).*
- *Analyzing the available solar-radiation databases for high-latitude locations (Paper III).*
- *Developing a regression model to improve the analyzed datasets (Paper IV).*

1.2 Overview of the thesis and appended papers

This thesis is structured in the following manner. Section 2 provides a general background of solar energy from a global and Norwegian perspective. Then, Section 3 explains basic Earth-Sun astronomical relationships that were used in the research and gives an overview of available solar radiation estimation technologies and resources. Section 4 explains the datasets used in the research, quality controls applied and validation metrics used to assess the datasets. Section 5 gives an overview of the previous research carried out on the estimation of solar radiation and presents the available knowledge gaps that this thesis aims to address. Section 6 presents the results from the research carried out. Finally, Section 7 provides a discussion on the results and future activities.

This thesis is composed of four papers that deal with the estimated solar radiation at high latitude locations. The results of this thesis are drawn from the appended papers, which are briefly presented below:

- *Paper I* presents a model that is based on the Hargreaves and Samani's maximum- and minimum-temperature difference model. In the proposed model, relative humidity was also used. The model was implemented and tested on eight locations in Norway for 10 years of data. Like other temperature difference models, this model had two distinct coefficients; one for coastal regions and another for inland regions. The proposed model slightly improved the Hargreaves and Samani model that it is based on. Some shortcomings of this model include having a highest temporal resolution of daily averages and inaccuracies introduced by having large temperature differences in clears sky-days. Importantly, as this model required *in-situ* measurements of temperatures and

relative humidity, its spatial resolution was limited to the locations where these meteorological variables are measured.

- *Paper II* presents a comparative analysis of CLARA-A1 and CLARA-A2 datasets for high latitude regions. The CLARA datasets are published by CM-SAF and these are constructed by using AVHRR instruments on-board the polar orbiting satellites. It was earlier found by some studies that satellite methods have high errors on snow-covered surfaces, which are frequent in high latitude regions. Because of this reason, CLARA datasets do not provide coverage on snow-covered regions. In this study, it was found that CLARA-A2 has less number of missing points than CLARA-A1. However, the new data points that were not available in CLARA-A1 had very high errors. Overall, it was found that CLARA-A2 is an improved data set, but it should be properly evaluated before using in regions that receive frequent snow cover.
- *Paper III* In this study, four dataset are compared and assessed for high latitude locations. Two of these datasets, CLARA-A2 and SARAH-2 are based on satellite models while the other two are reanalyses; a global reanalysis ERA5 and a regional reanalysis ASRv2. In this study, it was found that at location above 65ºN, CLARA-A2 provided better estimates then other datasets while below 65ºN SARAH-2 provided better estimates. It should be noted that SARAH-2 does not provide data above 65ºN. However, it was observed that for monthly averages of solar radiation, ERA5 provided comparable quality of estimates to CLARA-A2 and SARAH. ASR had the highest errors at all locations in this study. Furthermore, the cloud placement accuracy of ERA5 was analyzed and it was found that these errors are possibly due to overestimation of TCWC (total cloud water content) in intermediate-cloudy and overcast categories and an underestimation in clear-sky category. Nevertheless, ERA5 reanalysis can be used as a substitute to satellite databases for gap-filling procedures as the satellite datasets have missing values.
- *Paper IV* In this paper the knowledge gained from the previous papers is used to construct a novel data set by using an advanced regression method. In the previous studies, it was seen that generally satellite datasets underestimate solar radiations while reanalyses overestimate it. The hypothesis for this work is that combining two dataset with a regression model, where one dataset is having underestimation (Satellite based dataset) and other having overestimation (Reanalysis) can improve the estimated surface solar radiation. Random forest regression method was used with surface solar radiation estimates from ERA5 and CLARA-A2 for 31 locations in Norway and 16

years of data. In addition to surface solar radiation, solar azimuth angle, latitude, altitude and clear-sky index were used in the regression. The proposed dataset was improved on averages of daily, monthly, seasonal, and different-sky conditions. The regression model was tested on five locations from Sweden, which were not used in the training of the regression model. Almost the same degree of improvements was observed in Swedish locations as compared to the Norwegian locations that were used in the training.

2. Background

This chapter presents the background of the research undertaken in this thesis. Section 2.1 presents a historical overview on energy and climate change. In Section 2.2, the global energy demand and the available infrastructure are discussed. Then Section 2.3 gives an overview of the Norwegian energy infrastructure. Section 2.4 presents a global perspective on solar energy. Finally, in Section 2.5, the global solar energy perspective of Norway and current situation regarding solar installations are analyzed.

2.1 Historical overview of energy and climate change

Energy has played a central role in the evolution and prosperity of human societies. One of the first milestones of human evolution was the discovery of fire. This can be considered as the starting point of using energy for converting materials from one form to another, as in cooking food, refining metals or making pottery (1). Around 2500 years ago, humans started using energy from wind and water by inventing mills that convert energy from these sources to a rotary motion. One of the first documented evidences of using windmills was in Persia in the tenth century (2). This invention made it possible to grind edibles and produce other valuable resources. These pre-industrial advancements required a modest supply of energy, which was in turn restricted by the population growth and land availability (3). Apparently, the pre-industrial era can be considered as a hundred percent renewable based system, in which biomass, water and wind sources were the main drivers. This can be seen by observing the historical temperature anomalies in Figure 2.1, which shows a gradual increase in global temperatures after the industrial revolution.

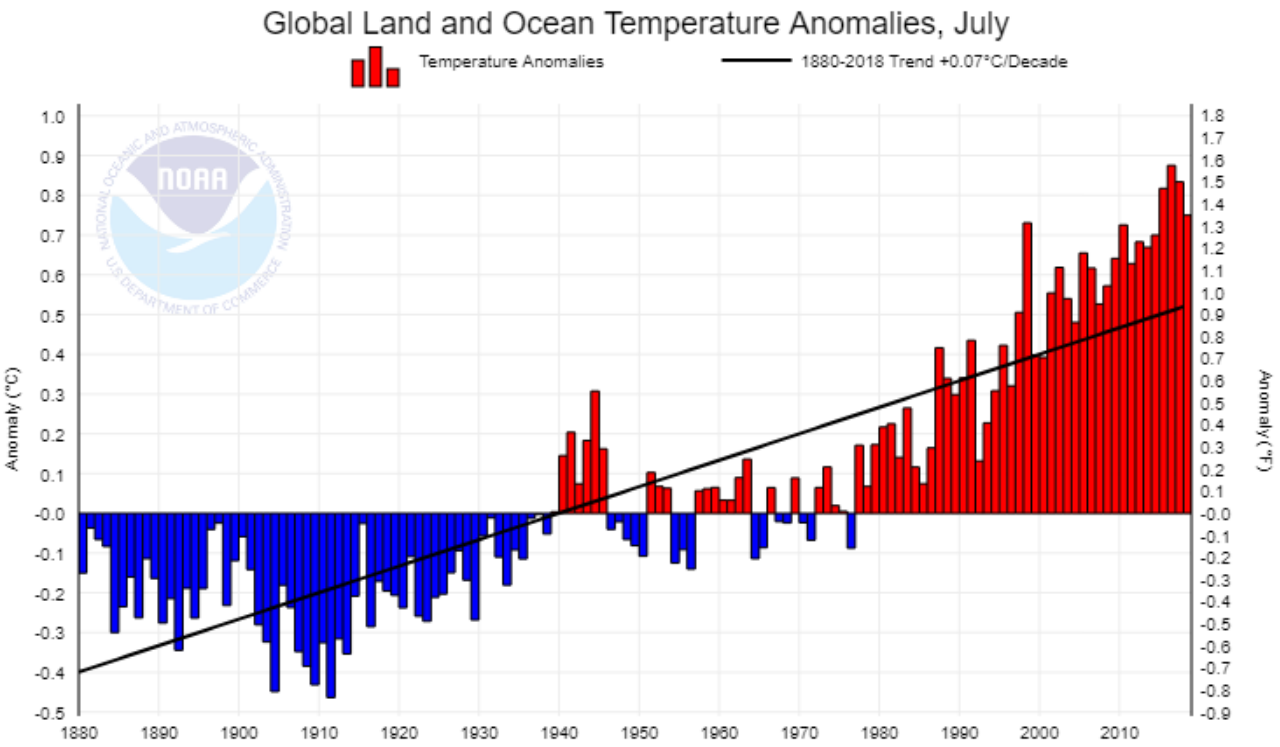


Figure 2.1: Temperature anomalies for 1880 to 2018 with respect to 20th century average. In this period, there is a positive trend of 0.07º C per decade. A sharp rise can be observed after the industrial revolution* (4).

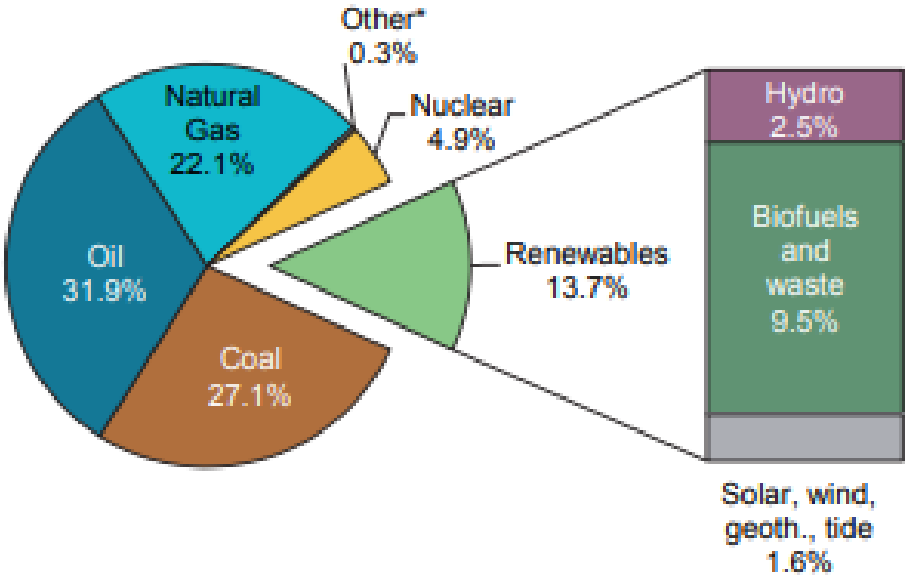
This period was followed by the Industrial revolution in Britain from 1760 to 1830. Industrial revolution brought major transformation in the socio-economic aspects, which on one hand brought an evolution in the living standards, but on the other hand, came with an increase in the emitted greenhouse gases (GHG). The turning point of the industrial revolution was the invention of steam engine, which unlike the cleaner wind and water mills, used fossil fuels. One of the earliest evidences of global warming caused by GHG was pointed out by Prof. Svante Arrhenius in 1896 (5). From the start of the 20th century, many scientists believed that carbon dioxide is the main cause for the rise in global temperatures but these studies lacked a concrete evidence (6). However, in 1985 World Climate Program published a report that pointed out the temperature increase by using powerful computers and sophisticated climate models (7). Meanwhile in Antarctica, research teams from France and Soviet Union performed ice drills and showed that both temperatures and carbon dioxide concentration have increased in the past ice ages. These studies developed a relation between temperature rise and carbon dioxide concentration in the atmosphere, however, this cause and effect relation between GHG and global temperature may be different from today (8). Following these and other researches, in

**NOAA National Centers for Environmental Information – accessed 10.11.2018*

1988 World Meteorological Organization (WMO) and United Nations (UN) established Intergovernmental Panel on Climate Change (IPCC) that was tasked with publishing climate change reports.

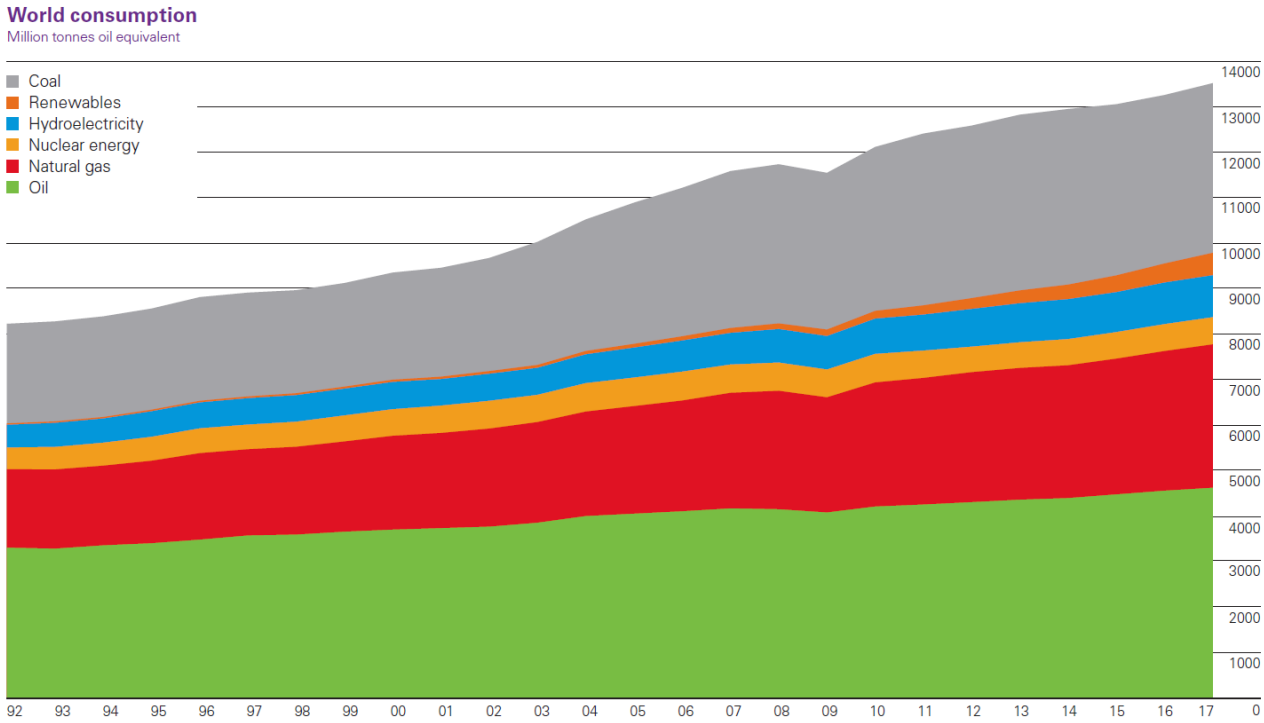
2.2 Current energy needs and infrastructures

In the history of human evolution, energy has played a major role. As humans evolved and progressed, there was a tremendous increase in the energy requirements of the world. The energy demands are still increasing every year and in 2016, the total primary energy supply (TPES) of the world was 13 761.4 Mtoe (million tons of oil equivalent) (9) . TPES is defined as the total supply of energy that is consumed locally. Figure 2.2 (a) shows an overview of the energy supplies for 2016. Moreover, 2017 saw an enormous increase in the global energy consumption, which grew at a rate of 2.2% as compared to 1.1% in 2016. Such an accelerated increase in the demands for energy consumption brings an increase in the emitted GHG. In the same period, the carbon emissions grew by 1.6% (10). The main sources of GHG emission are associated with production of electricity, heating and transport, which accounted for 49% of the total emissions in 2017. In the meantime, renewable energy share grew by 17%, higher than the last 10 years average (10). Even though there was an increase in the share of renewable energy, a decrease in the GHG emission was not observed because of the increase in total energy demands, which is illustrated in Figure 2.2 (b).



(a)*

*Based on IEA data from Renewables information: overview © OECD/IEA [IEA 2018], www.iea.org/statistics, License: www.iea.org/t&c



(b)

Figure 2.2: (a) The total primary energy supply (TPES) for 2016. The major portion of the energy supply is from fossil-fuel based systems, while there is a constant increase in the renewable energy sources. (b) The increase in world consumption in terms of different resources from 1992 to 2017* (11).

2.3 Energy Overview of Norway

The Norwegian energy supply has one of the highest share of renewable energy in the world. Hydropower provides the backbone for the energy infrastructure in Norway, providing 96% of the electricity and a large reservoir capacity of 85 TWh (half of the total in Europe). Other renewables account for a mere 2% of the generated electricity. Among the IEA member countries, Norway has the fifth lowest share of fossil fuels in TPES, although this has increased by 10% in the past ten years. Oil is one of the biggest industries in Norway and it has enabled Norway not only to be independent from energy imports but also made it one of the major exporters of energy. However, Norway has to rely on importing electricity periodically from the Nordic market to meet its peak demands. Norway has a unique energy overview; on one

**IEA/IRENA Global Renewable Energy Policies and Measures Database © OECD/IEA and IRENA, [28.11.2018]*

hand most of the energy generated in the country comes from hydro power, and on the other hand Norway is Europe's largest exporters of Oil (9). Figure 2.3 shows the overview of energy production of Norway.

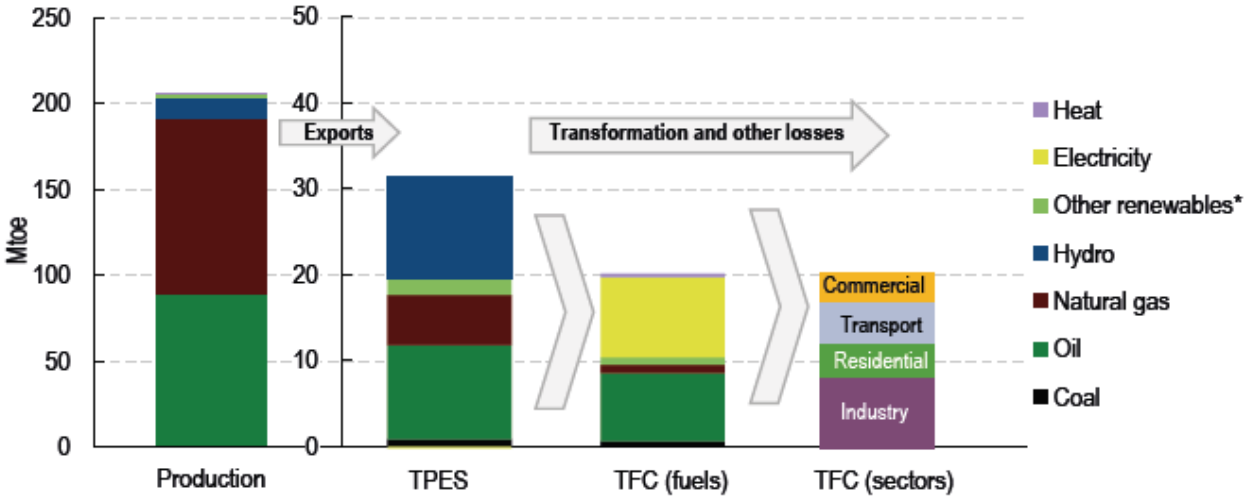


Figure 2.3: Energy production overview of Norway. TPES represent the total primary energy supply, which is defined as the total supply of energy that is consumed locally, expressed in million tons of oil equivalent. Total final consumption (TFC) represents the final consumption by the end user in the form of electricity, heat, gas, oil etc.* (9).

The total final consumption (TFC), which is defined as the final consumption by the end user in the form of electricity, heat, gas, oil etc., has been around 20 Mtoe over the past 15 years. As depicted in Figure 2.3, industry is the largest energy-consuming sector with 40% of the TFC share. This is followed by transport, which accounts for 24% of TFC. Norway has been very progressive towards climate change mitigation and sustainability, and in this regard, the government plans to reduce emissions by 30% from 1990 to 2020, and become carbon neutral by 2030. By 2050, the state targets include to become a low emission society. Although Norway still has large shares of hydropower in the electricity mix, the oil industry and transport sector use fossil fuels that contribute the most to the carbon emissions. One of the primary targets to become a low emission society would be to use renewable sources in these sectors. The transport sector is very progressive where the government has implemented strong incentives for electric vehicles (9).

**IEA/IRENA Global Renewable Energy Policies and Measures Database. Energy Policies of IEA Countries: Norway 2017 Review*

2.4 A Global perspective on Solar Energy

The Paris Agreement signed on December 2015, limits countries intent to the global warming to below 2ºC. To reach this target, solar energy will be one of the most important resources. Existing fossil-based energy systems can be replaced by more cleaner solar energy systems, meanwhile future energy needs can be fulfilled by using solar and other renewables. In 2016, renewable energy accounted for 18.2% of global TFC (10.4% of these systems were modern renewable, including wind turbine, solar photovoltaic (PV) etc.). A record increase in the installed PV capacity was observed in 2017 with 98 GW of PV additions, almost twice of the wind power additions and more net capacity than coal, natural gas, and nuclear power combined. The total global capacity of solar based energy systems reached 402 GW by the end of 2017 (12). These increments in installed capacity are largely due to the subsidies provided by the governments and the declining prices of PV.

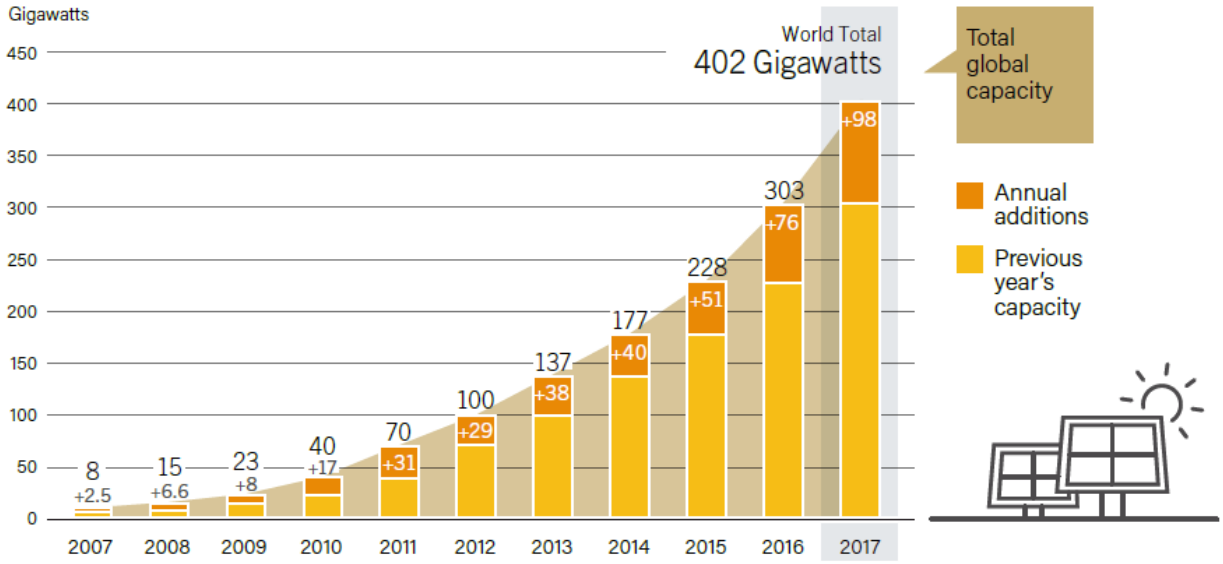


Figure 2.4: The increase in the installed PV capacity in the world from 2007 to 2017. 2017 saw a record addition of 98 GW and total installed capacity reaching 402 GW (12).

Even though there has been an exponential rise in PV and other renewable sources in the world, the demand for energy has also been increasing. To mitigate the effects of increasing energy requirements, and the consequent increase in the carbon emissions, renewable sources needs to increase at least six times faster if the goals set by the Paris Agreement are to be met (13).

**REN21. 2018. Global Status Report. (Paris: REN21)*

2.5 Solar energy in Norway

There is a common misconception about the feasibility of harvesting solar energy in the Nordic regions. Unlike equatorial regions that have a daily regular variation in received solar radiation, high latitude locations have a very different variation; as in these regions, midnight Sun occurs in the summer months with 24 hours of sunlight and polar nights occur in the winter when the Sun remains below the horizon. Because of these characteristics, the distribution of solar radiation is skewed towards the summer months. In high latitude regions, solar energy-based systems become viable only in conjunction with other sources that can provide back up in winter months. In the summer months, the Sun lie's above the horizon for a long time but PV systems become feasible only with at least one axis tracking. By employing a tracking system, the annual solar energy yields in Norway are comparable to that of Germany, which is the industry leader in PV installation.

In Norway, the penetration of solar PV or thermal has not been very large but recent years saw an exponential rise in the installed solar PV systems. Figure 2.5 shows the increase in the installed PV capacity from 2012 to 2017 in Norway.

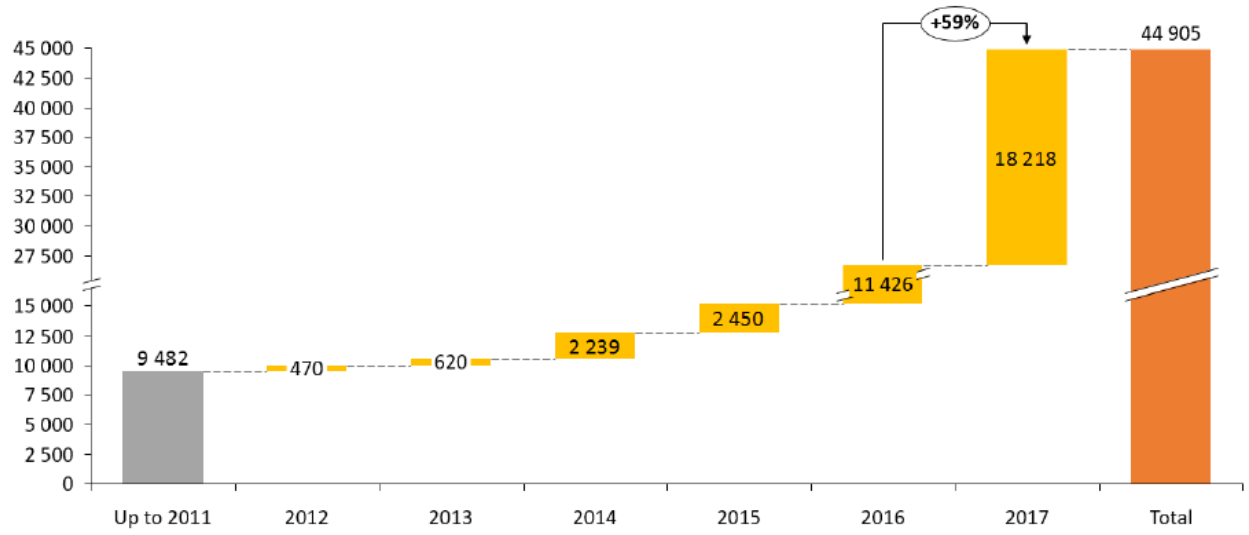


Figure 2.5: Installed capacity of solar PV in Norway. The growth in the PV installed capacity have been exponential in Norway with 2017 having the highest growth* (14).

The recent increase in the installed capacities of solar PV systems in Norway is also substantially due to the subsidies provided by government and the declining costs of these systems. A determining factor for the relative slow growth of solar systems is the price of

*IEA Photovoltaic Power Systems Programme

electricity in Norway, which is considerably lower than in central and southern Europe. A low electricity price means that it is harder for solar systems to compete.

3. Solar radiation

The technical and economic performance of solar thermal or solar PV systems depends on the total amount of solar radiation received on their exposed surface. Such estimations for specific locations give an insight into the pre-feasibility of these energy systems. This section describes the Earth-Sun astronomical relationships from the perspective of harnessing solar energy. Section 3.1 provides an overview of the potential of solar energy on the surface of Earth and different astronomical variables that affects it. Section 3.2 illustrates the path of the Sun for high latitude locations and demonstrates the usefulness of optimal angles and tracking strategies that increase the energy generation from solar energy systems. In the end, Section 3.3 gives an overview on the available solar radiation databases for Norwegian locations.

3.1 Harnessing energy from the Sun

The amount of energy from the Sun striking the surface of the Earth is very large. About 1.75 x 10^5 TW of solar power constantly strikes the Earth's surface. Even after considering a 40% loss from atmospheric cloud cover at any time, 1.05 x 10^5 TW is available on Earth's surface at any time. By using only 1% of the surface of Earth and converting it with a 20% efficiency, it would provide a resource base of 210 TW. The total global energy needs for 2050 are projected to be approximately 25-30 TW (15). These figures show that with a little effort most of the future energy demands could be met by using a clean and GHG emission free resource. However, there are a few hindrances in achieving such goals. Despite the fact that solar resource is abundant, one of its limitations is that it has a low flux density, which requires very large areas to be used as collectors. The Earth has a surface area of 510 million km^2, 1% of this surface is still a gigantic area. The second barrier is that most of the radiation falls on remote locations, which are far away from the human settlements. Equatorial regions between 25ºN and 25ºS receive large amounts of solar radiation on horizontal planes but most of these areas are desert regions (15). High temperatures, dust, lower availability of water and low population make these areas unfavorable for large installations. Some form of transmission infrastructure, which is expensive, must be developed prior to large installations in these areas (15). A viable solution can be achieved by installing medium and large power plants along with residential rooftop systems. Extreme northern and southern areas also receive adequate amount of radiation, but the average sun light duration is not constant throughout the year, as in equatorial regions. In these regions, tracking systems can enhance the generated energy production. The third deterrent is the need for storage. Solar energy is intermittent in nature and to be effective,

it needs a storage system that can provide backup when the Sun is below the horizon. Alternative methods are being developed that propose hybrid systems that employ solar, wind, hydro, biomass, and energy storage to flat out the intermittency (15).

3.1.1 Extraterrestrial radiation

The Sun emits tremendous amounts of energy while maintaining a surface temperature of 5760 K. To sustain all kinds of life, the Earth uses this energy in various forms, *e.g.* photosynthesis, wind circulation, water circulation, vitamin D, and so forth. Sun emits its energy in the form of electromagnetic radiation mostly in the range of 0.15 μm to 120 μm. This bandwidth covers visible spectrum in addition to ultraviolet and a part of infrared spectrums. The solar radiation received just outside the Earth's atmosphere is called extraterrestrial radiation (16). The value of extraterrestrial radiation changes throughout the year because of the changing distance between Sun and Earth (5.9% variation over a year). The variation in distance occurs because the Earth makes an elliptical orbit around the Sun. In solar radiation studies, a constant value of extraterrestrial radiation that is averaged over a year, called solar constant, is taken as 1361.1 Wm^{-2} (17). Figure 3.1 shows the daily average extraterrestrial irradiance on a horizontal surface for Tromsø, Norway over a year (adapted from Paper I).

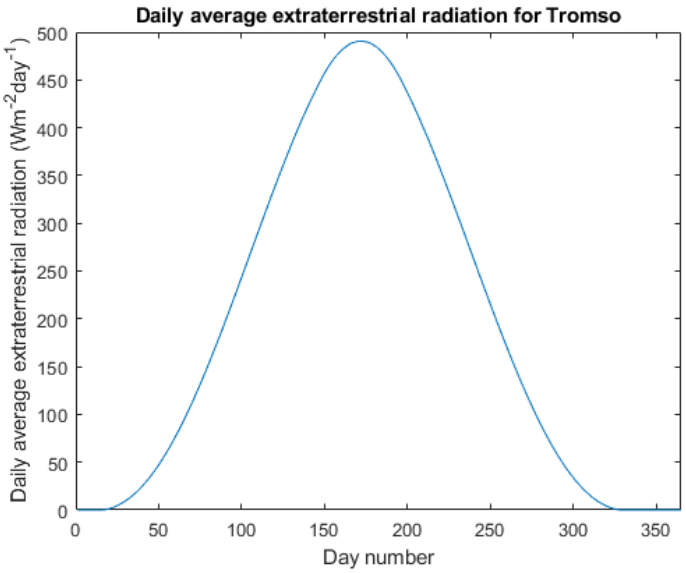


Figure 3.1: Daily average extraterrestrial radiation for Tromsø. The extraterrestrial radiation is zero in winter months because the Sun remains below the horizon (adapted from Paper I).

3.1.2 Solar radiation at the surface of Earth

To reach the surface of the Earth, the extraterrestrial radiation travels through the atmosphere. About 30% of the extraterrestrial radiation is reflected back by the atmosphere and 16% is absorbed by atmospheric gases (16). While passing through the atmosphere, solar radiation interacts with atmospheric gases like carbon dioxide, ozone and water vapors that cause absorption and scattering at certain wavelengths. Figure 3.2 depicts the spectral distribution of solar radiation outside the atmosphere, on the surface of Earth and the absorption caused by the atmospheric gases. Table 3.1 lists the distribution of energy in the solar spectrum on the surface of the Earth. It can be seen from Figure 3.2 and Table 3.1 that most of the energy in the terrestrial solar radiation lies in the visible and infrared bandwidths.

Table 3.1: Distribution of spectral contents of the Sun on the surface of Earth (16)

Type of radiation	**Range of wavelengths (nm)**	**% of energy carried**
Ultraviolet	150 to 380	7.6
Visible	380 to 720	48.4
Infrared	720 to 4000	43
Other	>4000	1

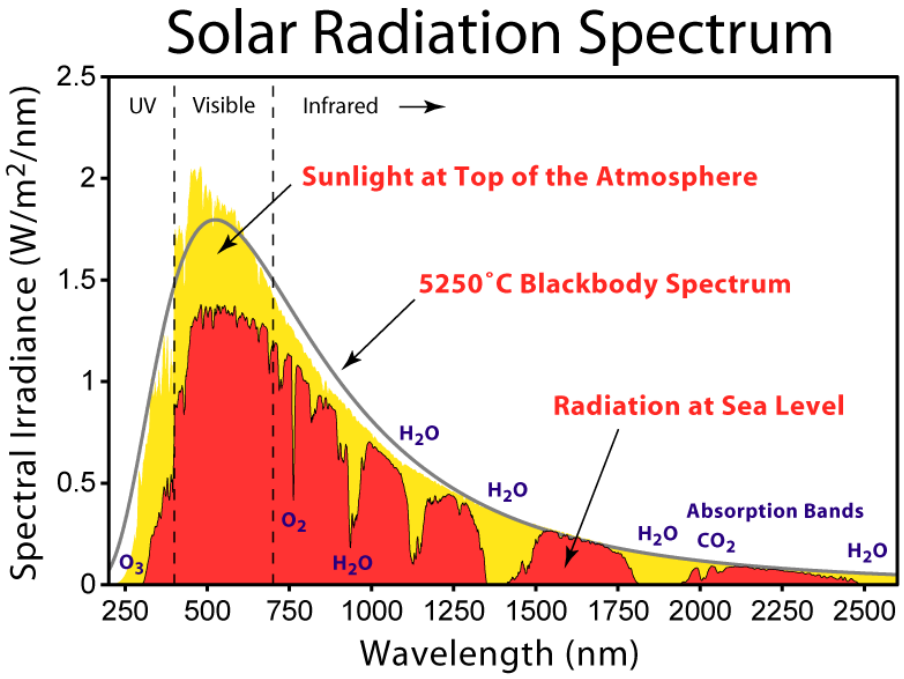


Figure 3.2: The spectral irradiance of the Sun is shown for extraterrestrial and terrestrial radiation. The absorption caused by different atmospheric gases is also indicated (18).

**Creative commons, Solar Spectrum, by Robert A. Rohde as part of the Global Warming Art Project*

When the solar radiation passes through the Earth's atmosphere or airmass, the direct optical path length that sunlight travel through the atmosphere determines the attenuation caused by scattering or absorption by the atmosphere (16). Airmass can be calculated by Equation 1:

$$AM = \frac{1}{\cos(\theta)}, \tag{1}$$

where θ is the angle that rays of the Sun make with the vertical at any point on the surface of the Earth. The radiation outside the atmosphere (extraterrestrial radiation) is referred to as AM0. In equatorial or tropical regions, the Sun is at the highest position at solar-noon and the solar radiation has to travel the least amount of distance to reach the surface. This is type of airmass is called AM1. However, at high latitude locations, the elevation of the Sun remains very low and the solar radiation has to travel relatively longer through the atmosphere when compared to equatorial regions. θ, the angle the sunrays make with the vertical can be related to the solar elevation or altitude angle. It is the angular height of the Sun in the sky measured from the horizontal (19). Solar elevation is expressed by the following equation:

$$\alpha = 90 + \varphi - \delta, \tag{2}$$

where, α is the solar elevation, φ is the latitude of a location and δ is the declination angle (explained later). As this angle determines how much the sunlight has to travel in the atmosphere before striking the surface of the Earth, it plays a critical role in determining the total production from solar collectors and their optimal angles.

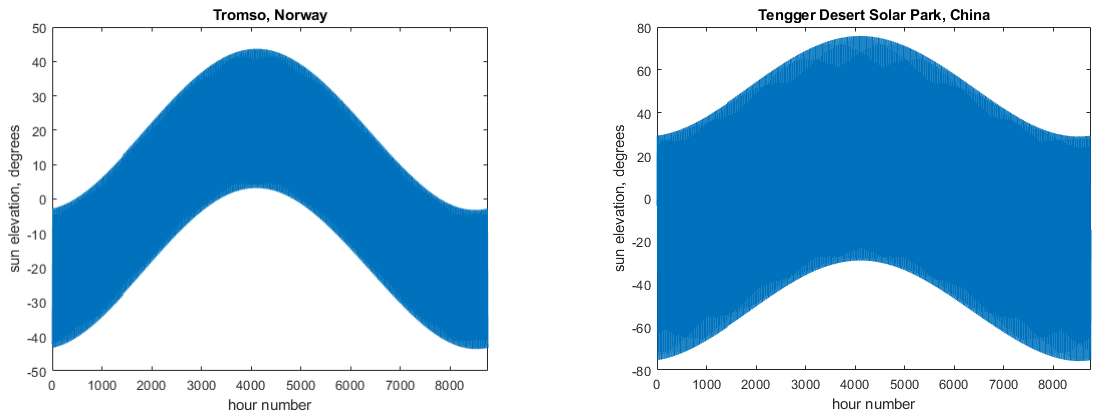


Figure 3.3: Solar elevations for Tromsø, Norway and Tengger Desert Solar Park, China (biggest solar photovoltaic installation in the world). Tromsø lies at 69ºN latitude while Tengger Solar Park lies at 38ºN latitude. It can be seen that the highest elevation on summer solstice is 43º in Tromsø while it is 76º at Tengger Solar Park. X-axis shows the number of hours in a year (8760 for non-leap years)

Figure 3.3 shows a comparison of solar elevations between Tromsø, Norway and Tengger Desert Solar Park (located in Zhongwei, Ningxia, China). Tengger Desert Solar Park is the largest solar PV installation in the world with a total peak power output of 1500 MW (20). The highest solar elevation occurs in summer solstice (21 June), which in Tromsø is low at 43º while at Tengger Desert Solar Park it is 76º. The negative values in Figure 3.3 show that the Sun is below the horizon. Another interesting point to note is even though the solar elevation is higher in Tengger Desert Solar Park, on 21st of June the sunsets while in Tromsø, even after having a low solar elevation, the Sun remains above the horizon. This indicates that despite having low solar elevation, the high latitude locations receive more solar radiation in summer months mainly because the Sun remains above the horizon for relatively longer periods.

3.1.3 Declination angle

The declination angle is defined as the angle between the equator and a line drawn from the center of the Earth to the center of the Sun (16, 19). Declination angle is independent of latitude and longitude, and it is responsible for changes in seasons. The maximum change in declination angle is less than 0.5º, which occurs at the equinoxes and for this reason a constant value is usually taken for a day (21). Declination angle is expressed by the following equation.

$$\delta = 23.34 \times sin\left[\frac{360}{365}(284 + n)\right], \quad (3)$$

where δ is the declination angle and n is the day number (from 1 to 365). Figure 3.4 illustrates a plot of declination over a year.

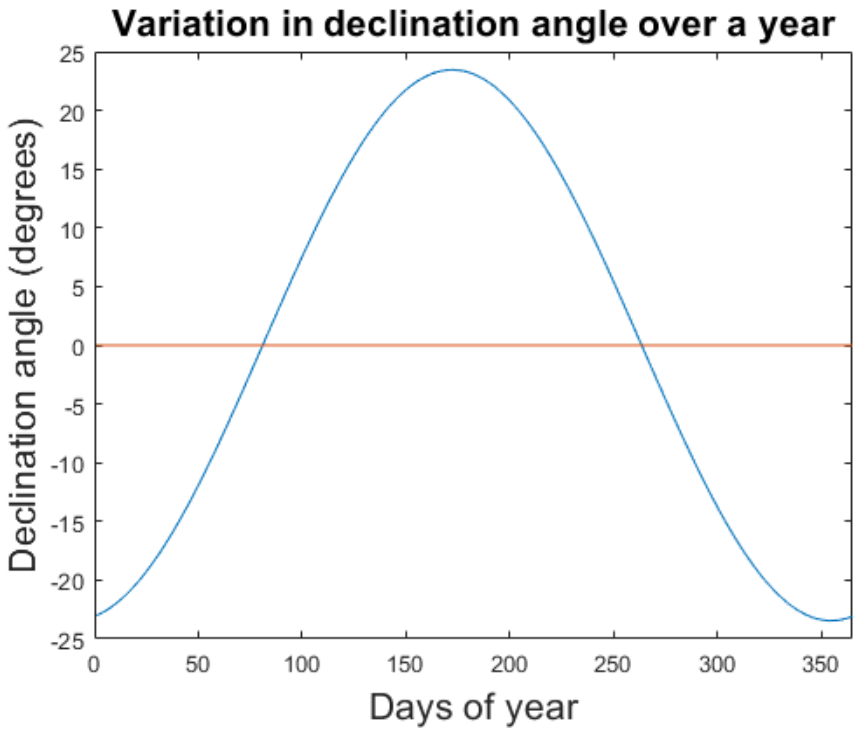


Figure 3.4: The variation in declination angle over a year is shown here. The declination angle changes from -23.45º (December solstice) to 23.45º (June solstice), while twice a year the value of declination angle becomes zero at equinoxes (16). The x-axis represents the day number of the year ranging from 1 to 365 for non-leap years.

3.1.4 Equation of time

A solar day, not necessarily 24 hours, is based on one full revolution of the Earth around its axis. The solar day varies in length throughout the year because the Earth sweeps unequal areas on the elliptic plane as it revolves around the Sun because the Earth's axis is tilted with respect to the elliptic plane (21). The inconsistency caused by such a revolution is called equation of time. As much as 16.45 minutes of variation can occur because of the eccentricity of Earth's orbit (19, 22). The equation of time is given by:

$$EoT = 9.87 \times \sin(2B) - 7.53 \times \cos(B) - 1.5 \times \sin(B), \quad (4)$$

where B is given by,

$$B = \frac{360}{365}(d - 81), \quad (5)$$

where d is the day number (from 1 to 365). The equation of time is shown graphically in Figure 3.5.

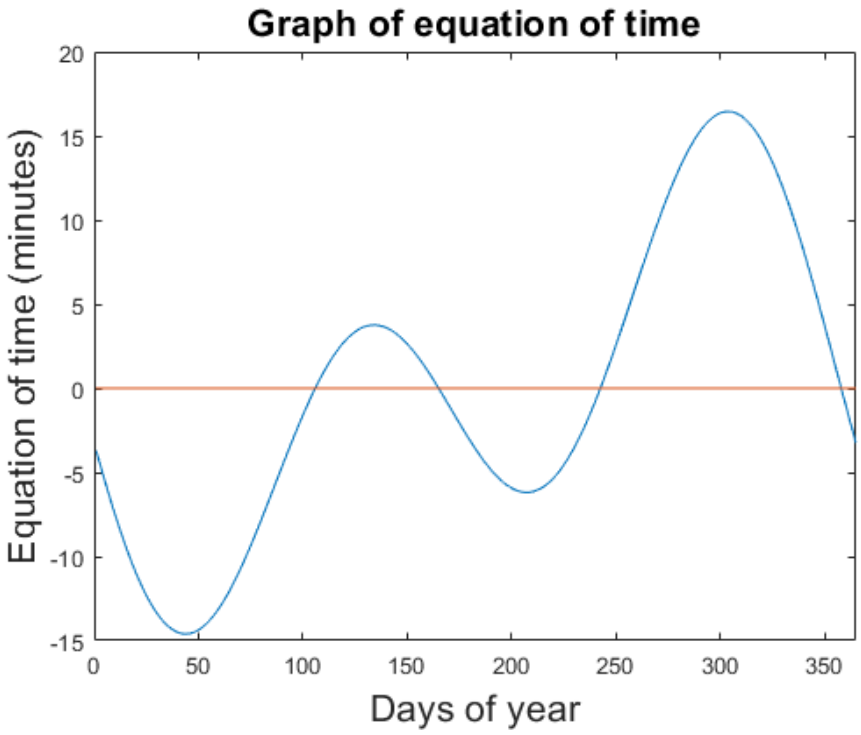


Figure 3.5: Graphical illustration of equation of time. The change in solar time occurs because of the eccentricity of Earth's orbit. A maximum of 16.45 minutes of variation occurs in a year. The x-axis represents the day number of the year ranging from 1 to 365 for non-leap years.

3.2 Path of the Sun at high latitude locations

The path of the Sun relative to an observer changes significantly with latitude. Figure 3.6 shows the path of the Sun in Tromsø for solstices (when the Sun is farthest away from the Earth on June 21 and December 21) and equinoxes (when the Sun is exactly above the equator on March 21 and September 23). For Tromsø, the path of the Sun for December 21 is not visible because the Sun lies below the horizon; however, on June 21 the Sun remains above the horizon for 24 hours, hence a 360º visibility of the Sun. Nevertheless, comparing this Sun path to the one shown for Gavdos (Greece), the southernmost point of Europe (34º50′N 24º05′E) in Figure 3.7, it can be seen that at lower latitude, optimally inclined solar collectors can be feasible while at higher latitudes, solar collectors with tracking systems can increase the output significantly.

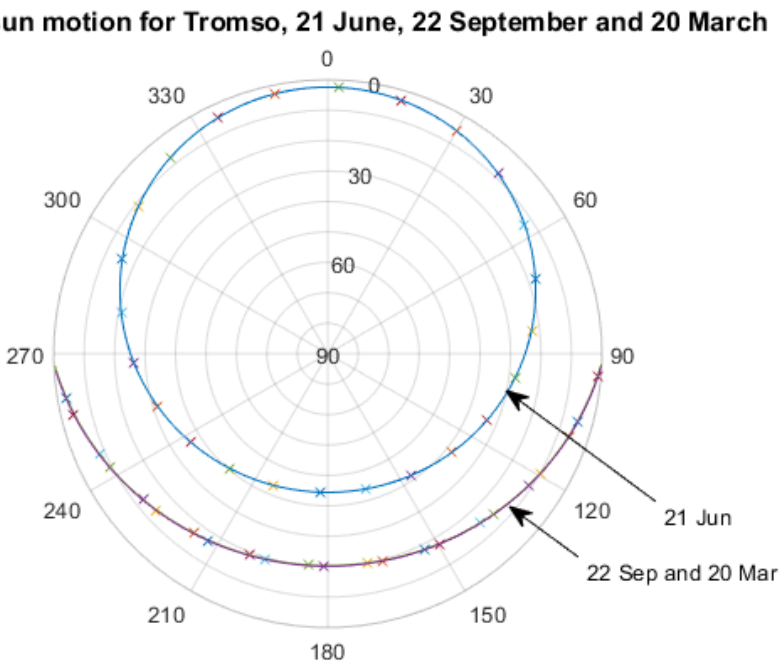


Figure 3.6: Path of the Sun motion for Tromsø. The paths are plotted for solstices and equinoxes, *i.e.* 21 June, 22 September, and 20 March. The path for 21 December is not visible because the Sun does not rise above the horizon.

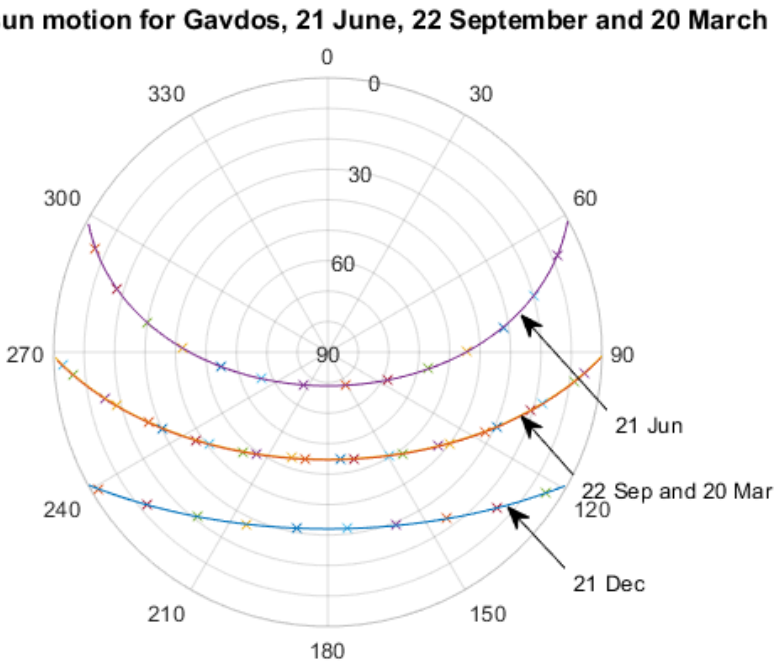


Figure 3.7: Path of Sun motion for Gavdos, Greece. The paths are plotted for solstices and equinoxes, i.e. 21 June, 22 September, 20 March, and 21 December.

3.2.1 Optimal angles for fixed collectors

The angle of incidence of sunlight on a solar collector changes with time of the day and day of the year, as shown in Figures 3.6 and 3.7. A solar collector will harness more energy if its surface is oriented towards the Sun at all times. In most cases, primarily due to economic reasons, solar collectors are installed with a fixed optimal tilt. A rule of thumb for the optimal tilt is shown by equation 6.

$$\beta = \phi - \delta, \tag{6}$$

where β is the optimal inclination angle in degrees, ϕ is the latitude and δ is the declination angle. Over a year, as the average of declination angle δ is zero, the optimal inclination angle for a year at a particular location would be equal to the latitude of that location. It can be seen from Figures 3.6 and 3.7 that the optimal surface azimuth angle for the northern hemisphere is true south. Although, on the basis of average declination angle, specific optimal tilts could be calculated for different months or seasons to optimize solar energy systems (16). By using an optimal tilt angle, the received solar radiation at the surface of the solar collector can be increased by 10 to 25% when compared to horizontally mounted collectors (22).

3.2.2 Solar energy systems with tracking

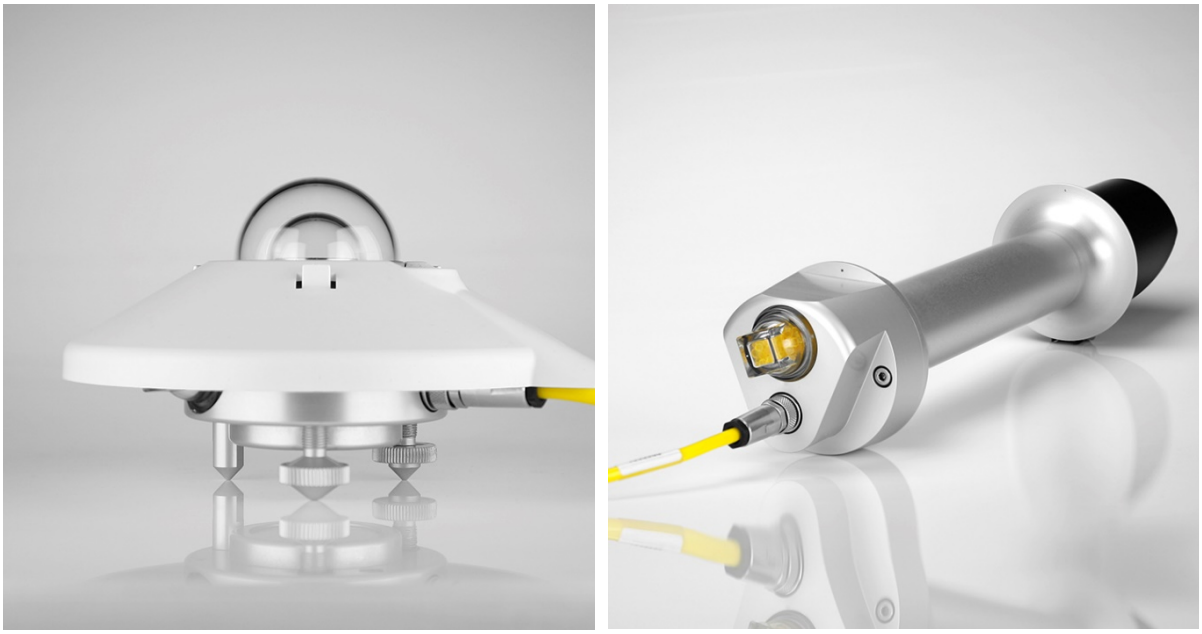
A solar collector mounted on a tracking system keeps the plane of the collector perpendicular to the incoming sunlight at all times. Such a tracking system increases the energy production by 30 to 50% when compared with stationary optimally inclined systems (23-26). In a two axis tracking system, the surface of the solar collector is always kept perpendicular to the incidence angle of the Sun. However, a single axis tracking system has one degree of freedom, which acts as axis of rotation. Usually, the axis of rotation in such a system is aligned along the true north meridian.

3.3 Estimation of surface solar radiation

This section gives an overview on the measurement and estimation of surface solar radiation. Section 3.3.1 explains the equipment used to record solar radiation at ground. This section also provides an overview of the available databases of solar radiation. In Section 3.3.2, the availability of ground measurements and solar radiation databases in Norway are explored.

3.3.1 Global solar resource estimation

The most accurate way to record solar radiation is by using equipment like pyranometers or pyrheliometers as shown in Figure 3.8. Pyranometers are used to measure global irradiance (or in most cases, global horizontal irradiance (GHI) as a pyranometer is installed on a horizontal plane). To record the direct normal irradiance (DNI), a pyrheliometer is used.



(a) Kipp and Zonen CMP11 pyranometer (b) Kipp and Zonen CHP1 pyrheliometer

Figure 3.8: Kipp and Zonen's CMP11 pyranometer and CHP1 pyrheliometer. Pyranometers are used to record global horizontal irradiance and pyrheliometers are used to record direct normal irradiance (27)*.

From the publicly available ground measurements of solar radiation in Norway, none of the stations provide DNI. Figure 3.9 depicts the available stations from Norwegian Institute of Bioeconomy Research (NIBIO) network that provide ground measured solar radiation data in Norway. It can be seen that most of the measurement stations are in the southern part of the country. The data from NIBIO is used in all the appended papers.

*Kipp and Zonen Instruments. Reprinted with permission.

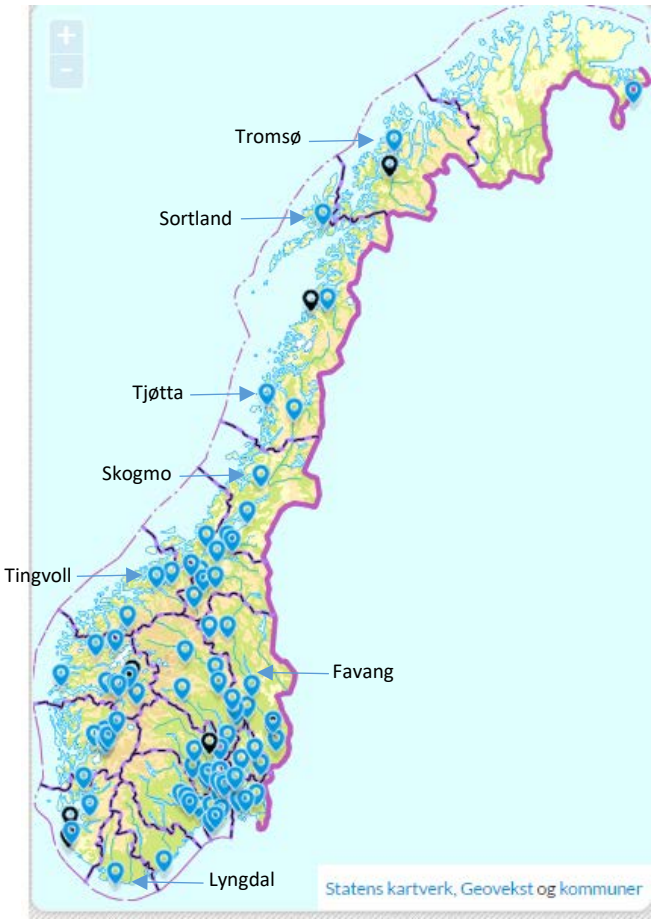


Figure 3.9: The NIBIO network provides ground-measured GHI data. Most of the stations in this network are in the southern part of Norway*.

Other indirect methods to estimate solar radiation explored in this thesis include satellites models, reanalyses and empirical models. Satellite models that are used to calculate solar radiation are well developed and widely used and provide solar radiation estimates with reasonable accuracy. Reanalysis, both global and regional, are also used to estimate solar radiation. Although these have lower accuracy than satellite models but very recent versions of reanalysis are improving and becoming sub-par with satellite estimations (28, 29). Generally, it is seen that satellite methods underestimate solar radiation while reanalysis overestimate (29-31). Empirical models exploit the relation between solar radiation and meteorological variables like sunshine duration, cloud cover, precipitation, humidity, temperature and so on. These models are considered as the least accurate (32).

Some specialized products provide solar radiation estimates by using the above-mentioned techniques. The PVGIS 4 (33), is one such product that provides solar radiation estimates based on CM-SAF Meteosat geostationary satellite images. The extent of the data provided by PVGIS

**Agrometeorology Norway, lmt.nibio.no. Reprinted with permission.*

is approximately 70ºN to 70ºS and 70ºW to 70ºE; however, the uncertainty in data is high at the edges of the coverage. The new version of this web database called PVGIS 5 is available for testing, for more information refer to Huld, Pascua (34).

Other products include S@tel-light, which provide solar radiation estimates for central and western Europe for the years 1996 to 2000 (35). Figure 3.10 presents an example of the coverage of S@tel-light for Norway. Another such database called SoDa (Solar radiation data), which is based on Helioclim 3, provides solar radiation estimates from Meteosat geostationary satellites. This database is also limited to -66º to +66º both in latitude and longitude (36). The data is available cost-free for a short time scale, while for longer time series there is an annual subscription. The SolarGIS is another such web application providing solar radiation estimations at 250m x 250m spatial resolution but this database is also limited to -60º to +60ºN and the data is available from 2004 onwards (37). Meteonorm is another such paid global database that is widely used. The data in Meteonorm covers the period from 1986 to 2005, with a total number of 1942 ground-measuring stations in the database. Meteonorm uses both ground measurements and geostationary satellite data to derive an interpolated global radiation dataset (38). Solem (39) is another such kind of a data set based on geostationary satellites (40). Most of the data sets based on satellite methods mentioned here are limited to 60º to 70ºN of latitude because they mostly use geostationary satellite that do not provide coverage above these limits. In addition, their accuracy becomes worse when moving towards high latitude regions.

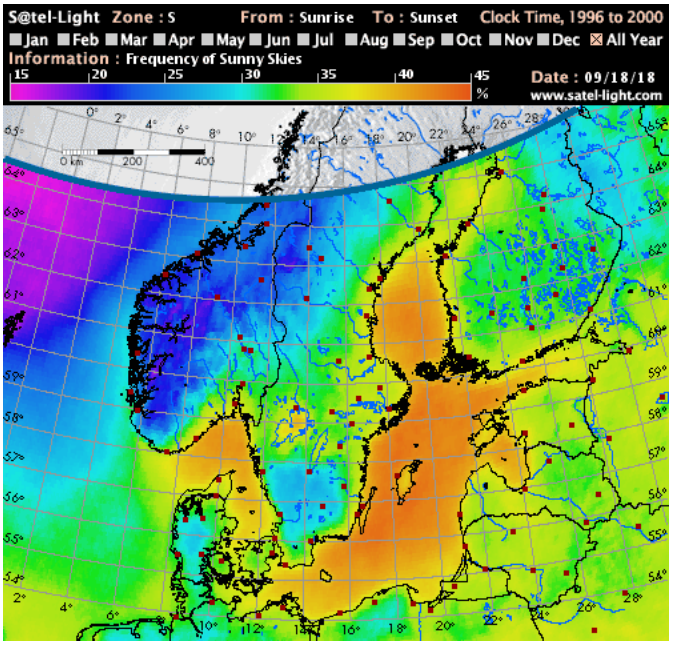


Figure 3.10: Map from S@tel-light showing the frequency of sunny skies. It can be seen that the data is limited to less than 65ºN* (41).

Likewise, a reanalysis product by Swedish Meteorological and Hydrological Institute (SMHI) called STRÅNG provides surface solar radiation estimates for Nordic regions with a grid of size 630 x 779. This product uses Mesan meteorological analysis model to produce the input and output fields (42). The input data for the product are derived from AROMIE numerical weather prediction system which is maintained at SMHI. This product provides instantaneous fields of global radiation, direct radiation and sunshine duration at a horizontal resolution of about 2.5 x 2.5 km and a temporal resolution of one hour. The accuracy of STRÅNG is approximately 30% for the global horizontal irradiance and 60% for the direct irradiance. Figure 3.11 shows a coverage map of STRÅNG for June 2016 (43).

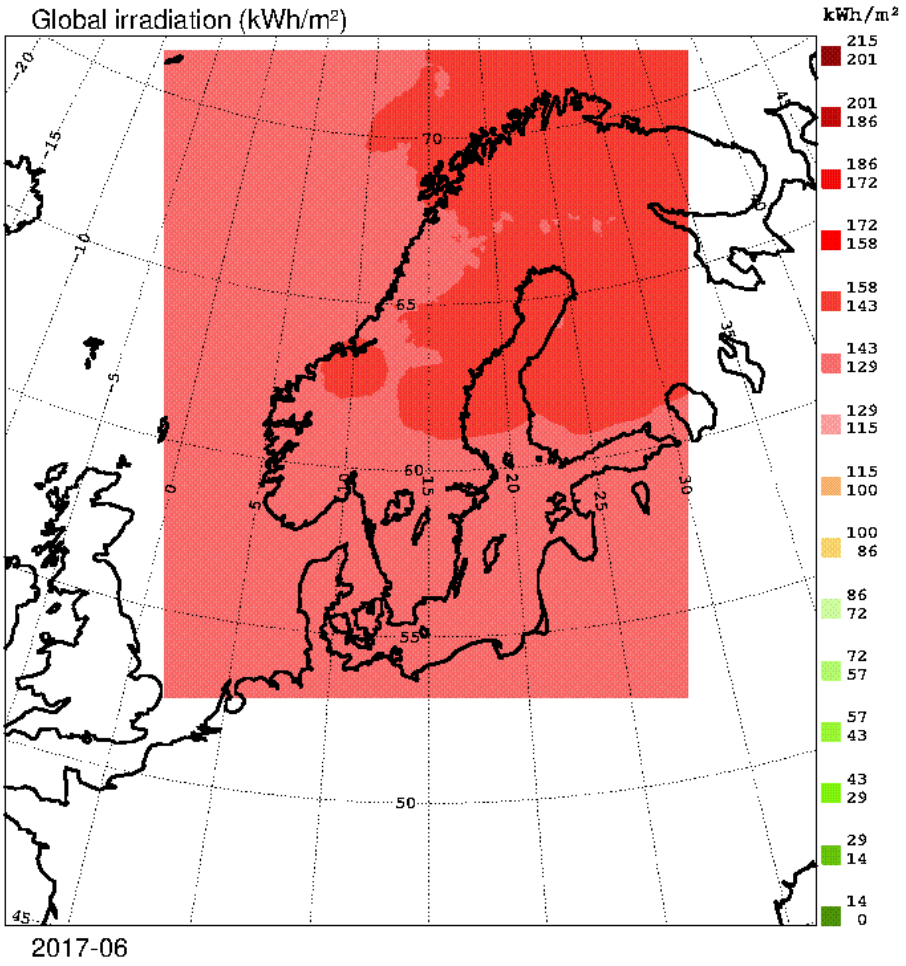


Figure 3.11: Global irradiation for June, 2017 from STRÅNG. This model covers the Nordic countries, the extent of the coverage can be observed from the figure (43).

3.3.2 Solar resource databases for Norway

Most of the satellite-based databases use geostationary satellites for a few reasons. First, these satellites have high spatial/temporal resolutions. Second, because of a large number of these

satellites, they provide coverage on almost all of the Earth. Third, a large population resides in the equatorial and mid latitude regions. However, these satellites do not provide coverage above 60º-65ºN. Most commercial and cost-free products of solar radiation use geostationary satellites in constructing their databases. Although some of these products provide coverage over southern parts of Norway, at region above 65ºN there are high errors in these datasets because of the slant viewing angles experienced by geostationary satellites. In addition to high latitudes, Norway presents a complex and challenging topography for estimating solar radiation. Figure 3.12 shows a digital elevation model of Norway (adapted from Paper II).

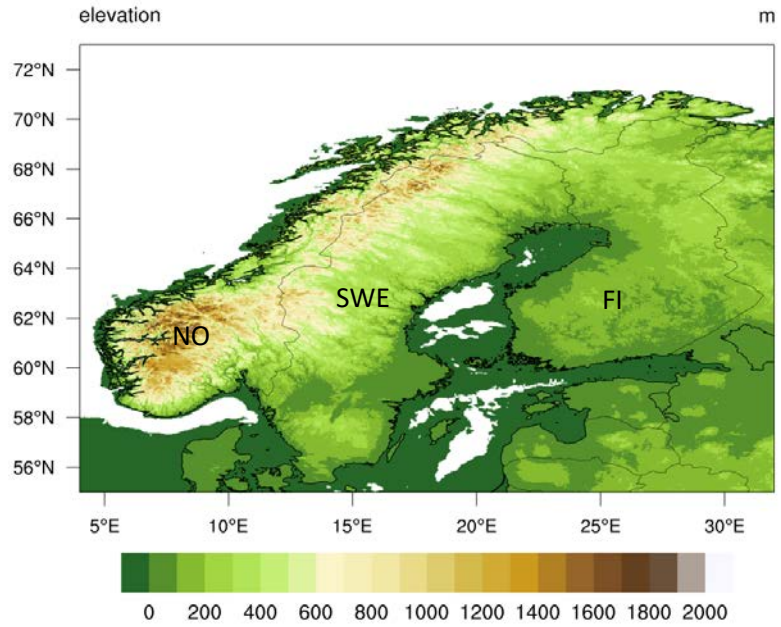


Figure 3.12: Digital elevation model of Norway, adapted from Paper II. It can be seen that there are complex elevations both in the southern and in the northern parts of the country.

Furthermore, Norway has a very low population density (2nd to Iceland in Europe). Because of a low population, there are only a small number of meteorological stations and even fewer of them record solar radiation. NIBIO is an agricultural network of pyranometers and it is the main agency responsible for maintaining and providing ground-measured solar radiation data for Norway. NIBIO has 47 stations in their database and at least 46 of these stations provide long-term solar radiation data series. The data is free to download as hourly, daily, and monthly means (44).

4. Methodology and data

In this chapter, the methodologies and the data used in this thesis and appended papers are presented. Section 4.1 gives an overview of the data used and their implications on the quality of estimations. Then Section 4.2 gives an overview of the ground-measured data. After this, Section 4.3 explains the different modelled data used in this thesis and appended papers. In Section 4.4, the quality control measures applied on ground-measured data are explained. Section 4.5 gives an overview of the regression method used in Paper IV. Section 4.6 shows the validation metrics used to evaluate the models. Finally, Section 4.7 shows the extraction methods used in the datasets.

4.1 Overview of the data

In this thesis, two types of data are used; ground-measured and modelled data. The ground-measured data include temperature, relative humidity and incoming shortwave solar radiation (temperature and relative humidity are used only in Paper I). The modelled data include solar radiation estimation from empirical model, satellite models, and reanalyses.

The most basic method to model solar radiation is by using empirical models. These models develop a relationship between incoming solar radiation and meteorological variables to estimate solar radiation. However, these models are site dependent and not as accurate as satellite models or reanalyses (32).

Reanalyses were first proposed in 1988 (45, 46) and are available since the mid-1990s. The global and regional reanalyses have been used to study both long- and short-wave down-welling solar radiation (47-53). Global reanalyses, as the name suggests, provide global coverage for major meteorological variables. Reanalyses are available at multi-decadal time scales and are usually cost-free. The data are available for monthly, daily, and sub-daily means (54-60). A regional-reanalysis is constructed by either dynamically or statistically downscaling a global-reanalysis. Weather research and forecast model (WRF) (61) is widely used in meteorology to downscale a number of global reanalysis under different configurations. One such example of a dataset, which is used in Paper II, is the Arctic System Reanalysis version 2 (ASR). ASR is a downscaling of ERA-Interim global reanalysis by using a polar optimized configuration of WRF (55). In Paper III and IV, ERA5 a global reanalysis from European Centre for Medium-Range Weather Forecast (ECMWF) is used.

Satellite models provide the most accurate remotely sensed estimates of solar radiation. Fritz, Rao (62) provided one of the earliest studies on the possibilities of estimating surface solar radiation by using visible sensors installed on satellites. They observed a high correlation between the radiance measured by the satellite sensors and ground-measured data. Later, Cano, Monget (63) introduced the basic idea that the surface solar radiation is inversely related to the top-of-atmosphere reflectance. Mainly, two types of satellites are used in these methods *i.e.* geostationary and polar orbiting satellites. Geostationary satellites are positioned at 35 786 km from the surface of the Earth and provide continuous observation on a spatial resolution of 3 - 5 km. However, geostationary satellites do not provide coverage in the polar regions because the apparent pixel size of the observation increases with latitude and longitude (64). For high latitude locations polar orbiting satellite are used as they provide coverage on poles. These satellites are positioned at around 800 km above the surface of the Earth and provide observation on a high resolution of 200 - 1000 m but with a low temporal frequency that varies with latitude (twice a day at equator and 14 times a day at the poles). The accuracy of geostationary satellite based datasets are better than polar orbiting satellite based datasets because of the high sensing frequency which takes into account the intermittent nature of solar radiation (32). Satellite methods generally underestimate down-welling shortwave solar radiation and reanalysis generally overestimate it (28, 29).

In the following subsections, the datasets used in this thesis and appended papers are explained in detail.

4.2 Ground-measured data

The ground-measured data used here was obtained from Norwegian Institute of Bioeconomy Research (NIBIO) for Norwegian locations, and Swedish Meteorological and Hydrological Institute (SMHI) for Swedish location. Both databases record average hourly measurement by Kipp and Zonen CMP11 or CMP13 pyranometers. The equipment is regularly maintained (on weekly or monthly basis) and datasets are quality controlled by the respective organizations (65, 66).

To evaluate the remotely sensed solar radiation estimates in different geographical conditions, the analyzed locations were divided into inland, coastal, above 65ºN and below 65ºN regions (Papers III and IV). The division between inland and coastal regions was established by observing the proximity of the stations to the shoreline. Regions within 30 km of the shoreline were considered as coastal. From the 31 Norwegian locations studied in Papers III and IV, 14

locations were classified as coastal and the rest as inland. The other two groups were made based on the latitude of locations where regions lying above 65ºN were grouped together while locations lying below 65ºN were put in another group. From the 31 Norwegian locations studied in Paper III and IV, 4 locations lie above 65ºN and 27 lie below 65ºN. For details on this classification, refer to the Appendix, Table A.

In Paper II, SMHI and NIBIO data were used and years having more than 10% of missing values were discarded. The rest of the years were having missing data and these were filled by using linear interpolation. In Paper III and IV, the ground-measured data was used after applying Baseline Surface Radiation Network (BSRN) recommended Long and Dutton quality control (67) and a quality control based on comparing the ground deviation with reanalysis and satellite model proposed by Urraca, Gracia-Amillo (68). These quality control procedures are explained in Section 4.4.

In addition to ground-measured solar radiation, temperature and relative humidity were used to construct a model to estimate solar radiation in Paper I. These data were acquired from NIBIO.

4.3 Model data

This section lists the model data used in this thesis and appended papers.

4.3.1 Empirical model based on maximum temperature difference and relative humidity

Empirical models estimate surface solar radiation by developing a relation between atmospheric transmissivity and other meteorological variables. One of the first such model was proposed by Ångström (69) in 1924. Ångström observed a high correlation between sunshine duration and daily solar radiation. Examples of other such empirical models use cloud cover (70), air temperature (71), precipitation and humidity (72, 73). However, the use of temperature and sunshine duration have been the most widely used technique in building such models because these variables are widely measured at weather stations (74, 75). In Paper I, a model based on Hargreaves, Samani (76) was proposed that uses the difference between maximum and minimum temperatures, and relative humidity in a day to estimate the average daily solar radiation. One of the shortcomings of empirical models is the use of difference between maximum and minimum temperatures in case of cloud-free conditions. In these conditions, the maximum and minimum temperature differences are relatively large due to low temperatures at night. In such cases the estimated solar radiation have high errors (32). However, the key

limitation of empirical model is the site-specific coefficient, which varies between coastal and inland regions (74), as these coefficients largely determines the accuracy of the estimated solar radiation (32). In addition, these types of models are dependent on ground based meteorological measurements and hence, these cannot produce a spatial distribution map of solar radiation estimates.

4.3.2 CM-SAF CLARA dataset

The Cloud, Albedo, Radiation (CLARA) dataset is a set of climate data records published by the Satellite Application Facility on Climate Monitoring (CM-SAF). The CM-SAF provides two categories of data: operational products and climate data records (CDR). The operational products are constructed by validating the data with on-ground stations and these are provided in near real time for variability studies in diurnal and seasonal time scales. However, CDRs are long-term data series that are used to assess inter-annual variability. CLARA-A1 and CLARA-A2 are two of such CDRs that provide long time series historical data. The CLARA datasets are based on polar orbiting satellites that provide a global coverage but their sensing frequency varies with latitude. These satellites have a sensing frequency of twice each day at the equator but with increasing latitude, the sensing frequency increases because of the overlap in the satellite swath. At the poles, these satellites have the highest sensing frequency of 14 observations each day. A single satellite has too low of a frequency to construct solar radiation datasets, hence, a series of satellites are used to obtain the surface solar radiation datasets.

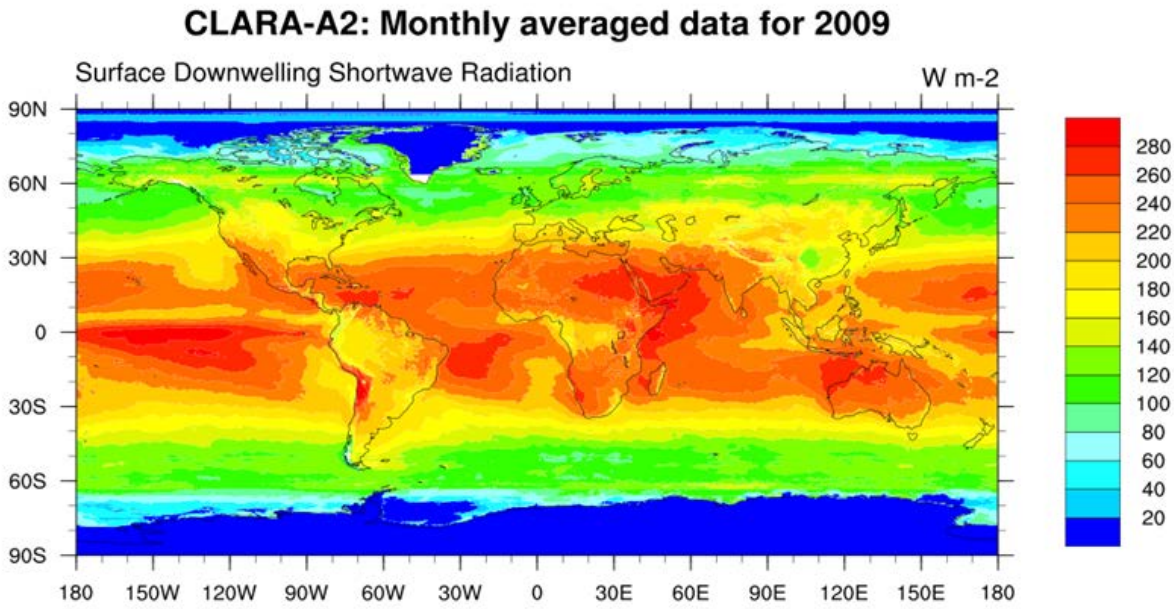


Figure 4.1: CLARA-A2 monthly mean solar radiation map for 2009 on a horizontal surface. From Paper II.

The first edition of this suite of dataset (CLARA-A1) was published in 2012 and it covers a 27 years period, from 1982 until 2009. This dataset consists of cloud, surface albedo and radiation budget products derived from the Advanced Very High Resolution Radiometer (AVHRR) sensors on-board the polar orbiting NOAA and Metop satellites (77). The second edition of this dataset, CLARA-A2, was released in December 2016. CLARA-A2 is available from 1 January 1982 to 31 December 2015, and constitutes an extension of 6 years relative to the CLARA-A1 dataset. Both of these datasets have global coverage with a spatial resolution of 0.25ºx0.25º on a regular latitude-longitude grid and provide daily and monthly averages of surface incoming shortwave radiation (SIS). To calculate daily averages, at least 20 observations of incoming solar radiation in each grid box are required; similarly, 20 valid daily observations are required to generate monthly averages (78). Along with SIS, CLARA also provides longwave up and down-welling surface radiation.

The fundamental method used in calculating surface solar irradiance from satellite observations is that the reflectance measured by the satellite instruments is related to the atmospheric transmittance. The SIS is calculated from the atmospheric transmittance (T) by the following equation.

$$SIS = E_0 \cos(\theta_0)\, T, \tag{7}$$

where E_0 is the extraterrestrial solar radiation and θ_0 is the solar zenith angle. The value of E_0 is set as 1368 Wm^{-2} in CLARA-A1, however, a revision in extraterrestrial radiation was performed by Gueymard (17) and the value of E_0 is set as 1361 Wm^{-2} in CLARA-A2.

In CLARA dataset, the transmittance is calculated from solar zenith angle, vertically-integrated water vapor, aerosol information and the cloud cover (obtained from AVHRR sensors). Finding solar zenith angle is straightforward and can be calculated accurately. The vertically-integrated water vapor and aerosol optical depth are not available in the AVHRR data and for these fields, external sources are used. For vertically-integrated water vapor, ERA-Interim Reanalysis (55) is used and the vertical ozone column is set to a constant value of 335 DU, as its variability has negligible impact on the estimated solar radiation. Aerosol information is taken from the modified version of the monthly mean aerosol fields from Global Aerosol Data Set/Optical Properties of Aerosols and Cloud (GADS/OPAC) climatology. In addition to this, the algorithm in CLARA also requires the surface albedo information. This is calculated based on spatial distribution of 20 surface types, which is obtained from the (SARB) Surface and Atmospheric Radiation Budget (part of the Cloud and Earth's Radiant Energy System (CERES)). In the algorithm, the cloud coverage is determined by using the visible channels of the AVHRR instrument. The first step in estimating surface solar radiation is the classification of the sky conditions. The Nowcasting SAF (SAFNWC) software is used to derive the information on cloud coverage for each pixel by using the information from the satellite sensors. If no cloud is detected (cloud free pixel), surface solar radiation is calculated by using only the auxiliary sources and clear-sky Mesoscale Atmospheric Global Irradiance Code (MAGIC) described in Haase, Calais (79) . If the pixel is classified as cloudy (cloud contaminated or fully cloudy), visible channels of the AVHRR instrument are used to derive broadband reflectance. This reflectance for each pixel is then transferred to broadband fluxes by using a bidirectional reflectance distribution function (BRDF). In the next step, these broadband top-of-the-atmosphere albedos are used to derive transmissivity through a look-up table approach. Finally, the transmissivity is used in calculating surface solar radiation, as shown in Equation 7 (80). In this dataset, all data points with solar zenith angles larger than 80º are set to missing values and solar zenith angles larger than 90º are set to zero. Because a temporally constant surface albedo is used in the algorithm, this dataset does not provide radiation estimates on snow and sea ice coverage areas because changes in the albedo of the snow-covered surfaces are not considered (81). For more information on the CLARA datasets and their accuracy, refer to Karlsson, Riihelä (77) and Karlsson, Anttila (81).

High-latitude locations like those studied here, may have a very different surface albedo than the temporally constant albedos considered in the algorithm. These critical points are identified by using the monthly mean CLARA-SAL (surface albedo) data record and the surface albedo used in the processing of SIS. All grid points with a difference in surface albedo exceeding 35% are masked out and set to missing data in final SIS record. This process introduces large number of missing data points in high latitude locations. Furthermore, the accuracy is reduced because at the available data points, a constant surface albedo is used which can vary from the real conditions. For this reason, the accuracy of the CLARA datasets in snow-covered areas is outside the target accuracy of CM-SAF. Further inaccuracies may be introduced by the miss-classification of SAFNWC software used in cloud detection. It was observed in Paper III that the aerosol information used in the CLARA dataset can introduce errors in clear-sky and intermediate-cloudy conditions because average monthly aerosol information can vary from the inter-annual and sub-monthly aerosol variability of a particular location (80).

In Paper II, CLARA A1 and A2 datasets are compared for Norwegian and Swedish locations. It was found that CLARA-A2, thanks to a new snow-detecting algorithm, has less number of missing values as compared to CLARA-A1. However, the new values that were not available in CLARA-A1 have large errors because these points mostly lie on the snow-covered surfaces. In Paper III, CLARA-A2 data set was evaluated and compared with SARAH, ERA5, and ASR. In this study, it was found that CLARA provides good estimates of surface solar radiation at location above 65ºN, where SARAH has no coverage. In Paper IV, this knowledge was used to construct a new dataset by using CLARA-A2 and ERA5 (explained in Section 4.3.4). The new dataset, which was constructed by using a random forest regression method (explained in Section 4.5), provides substantially more accurate results than CLARA-A2 and ERA5.

4.3.3 CM-SAF SARAH dataset

The second version of surface solar radiation dataset – Heliosat (SARAH-2) is a CDR of surface solar radiation by CM-SAF (82). The SARAH dataset covers a period of 31 years from 1983 to 2015 and the region from +65º to -65º in latitude and longitude. The spatial resolution of the data is 0.05ºx0.05º (approximately 5 km) and the data is available for 30 minutes instantaneous, hourly, daily, and monthly averages of surface incoming shortwave radiation on a horizontal surface (SIS), surface direct irradiance (SDI), sunshine duration (SDU) and effective cloud albedo (CAL), while spectrally resolved irradiance (SRI) is available as monthly means (83). To calculate daily averages at least three samples per day are required; similarly, 10 calculated daily observations are required to generate monthly averages.

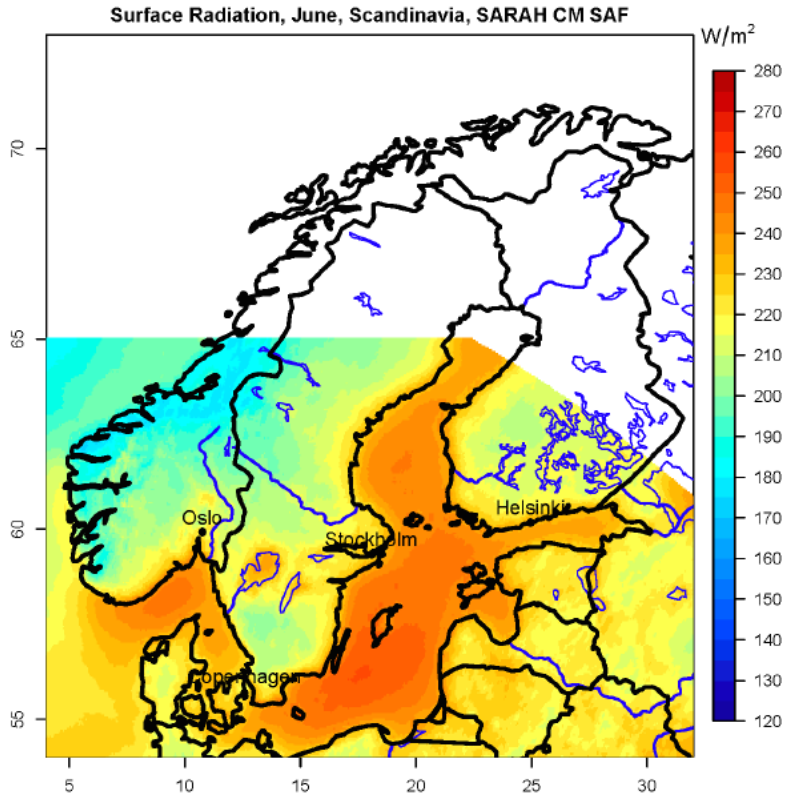


Figure 4.2: Mean surface solar radiation for 1983–2013 in June over Scandinavia from SARAH. The spatial limits for the dataset can be seen in the figure (84).

To obtain sufficiently large time series of data (spanning multiple decades), SARAH uses two generation of Meteosat satellites. The broadband visible channels from Meteosat Visible Infra-Red Imager (MVIRI) instrument on-board the Meteosat first-generation satellites (Meteosat-2 to Meteosat-7) and the Spinning Enhanced Visible and Infrared Imager (SEVIRI) instrument on-board the Meteosat second-generation satellites (Meteosat-8 to currently Meteosat-10) are used to calculate the shortwave surface irradiance.

The basic method of calculating surface solar radiation in SARAH is similar to that of CLARA. In SARAH, effective cloud albedo (CAL) and a clear-sky model are used to calculate surface solar radiation. The CAL is defined as the amount of reflected irradiance for all sky relative to the amount of reflected irradiance for clear-sky, and it is a measure of the cloud transmission and hence by calculating clear-sky radiation, the all sky radiation can be estimated. To calculate CAL, satellite data and a modified Heliosat method are used (85). This modification of the Heliosat method in combination with gnu-MAGIC/SPECMAGIC is called MAGICSOL. The Heliosat method uses reflection measurement given as normalized digital count to calculate the CAL. The effective cloud albedo from the Heliosat method is given by the following equation.

$$n = \frac{\rho - \rho_{srf}}{\rho_{max} - \rho_{srf}}, \quad (8)$$

where, ρ is the observed reflection for each pixel, ρ_{srf} is the clear-sky reflection and ρ_{max} is the estimated maximum reflectivity observed by the satellite sensor. The modifications made in this algorithm include a self-calibration algorithm that is based on an operational and automatic determination of the maximum reflectivity ρ_{max}.

In the next step, the clear-sky model and effective cloud albedo are used to calculate the surface solar radiation. The modified Heliosat method provides the broad band effective CAL but to consider the spectral effect of clouds a Radiative Transfer Model (libRadtran) based correction is applied. To calculate clear-sky radiation, SPECMAGIC model is used which is based on a so called hybrid eigenvector look-up table approach (86). The input parameters for gnu-MAGIC/SPECMAGIC are date, time, solar zenith angle, coordinates, effective cloud albedo (cloud index), water vapor column density, surface albedo, aerosol optical thickness, and single scatter albedo for aerosols. Monthly mean values of vertically integrated water vapor are taken from ERA-Interim global reanalysis record (55), and monthly mean aerosol information is taken from Monitoring Atmospheric Composition and Climate project (MACC) aerosol climatology. Surface solar radiation is derived from combining SPECMAGIC algorithm and effective cloud albedo (82). Improvements in the new version of the dataset (SARAH-2) includes the stability in the early years of dataset and during the change of instrument from MVIRI to SEVIRI in 2006 and correction of viewing geometry for slant viewing angles (87). For more information on the retrieval methods refer to Müller, Pfeifroth (88). SARAH-2 was used in Paper III to evaluate the solar radiation estimates in location below 65ºN.

4.3.4 ECMWF Reanalysis 5 (ERA5)

ECMWF Reanalysis 5 (ERA5), is the fifth generation of European Centre for Medium-Range Weather Forecasts (ECMWF) atmospheric reanalysis of the global climate and span a period of 1950 to near real time (39). At the time of writing, data from 2000 to 2017 is available. Further data back in time will be released in 2019-20 and will continue to update forward in real-time. In ERA5, the solar radiation variable has a spatial resolution of 31 km (0.28125ºx0.28125º) and an hourly temporal frequency. ERA5 uses the Integrated Forecasting System (IFS) cycle 41r2 with a state-of-the-art four-dimensional variational analysis (4DVAR) assimilation system. ERA5 has more pressure levels than ERA-Interim (the previous edition of ECMWF reanalysis) and more variables are made available for this reanalysis than for those of earlier generation. For more information on ERA5, refer to ECMWF (89).

In this study, shortwave surface downward radiation, shortwave surface downward radiation clear-sky, and total cloud water content (the vertically-integrated cloud water concentration) are used from this dataset. In ERA5, the incoming short wave radiation is obtained from a Radiative Transfer Model (RTM). This model simulates the attenuation in solar radiation caused by the atmosphere, therefore, the quality of estimated radiation depends on the RTM used. Reanalysis generally do not assimilate aerosol, clouds, or water vapor data, which increases the uncertainty in the estimated surface irradiance (49, 90). ERA5 was used in Papers III and IV.

4.3.5 Arctic System Reanalysis v2

In polar regions, it is difficult to determine current weather and climate trends from a long-term climatology perspective when compared to the rest of the globe, primarily because of limited number of meteorological stations (91). In these areas, reanalysis can be used as an alternative to provide such climatologies. To provide a long-term climatological data, the Arctic system Reanalysis was made available in 2010 (92). The second edition of this dataset was proposed in 2017 (93) called the Arctic system reanalysis version 2 (93). These are a set of regional reanalysis that are based on high-resolution regional assimilation of model output, observations and satellite data for the mid- and high-latitude regions of the northern hemisphere (94). In its core, ASR is a polar-optimized dynamic downscaling of ERA-Interim reanalysis by using Weather Research and Forecast Model (WRF) version 3.6.0 (95). The data set is available for the period of 2000 to 2012. The grid resolution is 15 km, which is finer than most global models and the previous release of ASR (ASRv01), whereas the time resolution of the dataset is 3 hours. The downscaling is optimized for Polar Regions, and polar physics is used where possible, including heat transfer through snow and ice, the fractional sea ice cover, the ability to specify variable sea ice thickness, snow depth on sea ice and sea ice albedo, as well as other optimizations including the Noah Land Surface Model. The area covered by this dataset is 1.2 x 10^8 km^2, which is about 50% of Northern hemisphere. Spectral nudging from ERA-Interim is applied on geopotential height, temperature, and wind components above 100 hPa on the inner domain. ASR uses three-dimensional variational analysis (3DVAR) for observations, including radiance data, from a number of satellites (93). Figure 4.3 shows the inner and outer domains used in ASR.

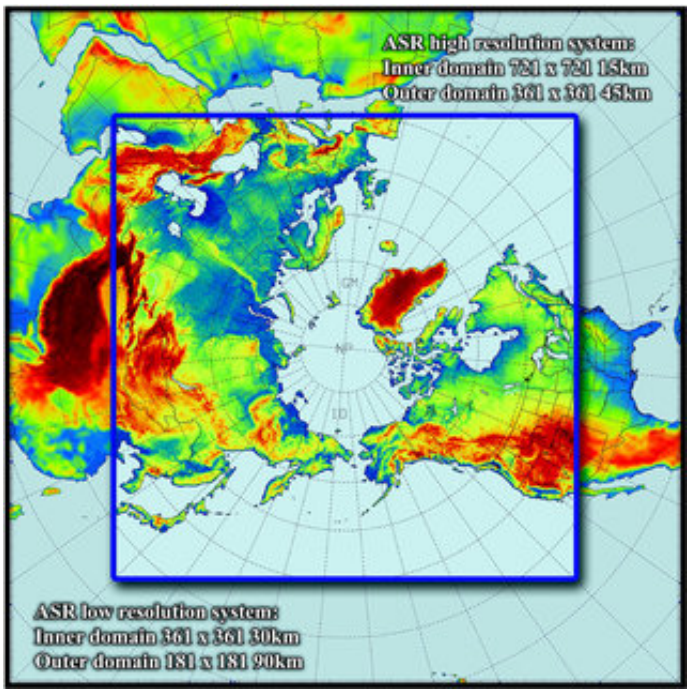


Figure 4.3: The inner and outer domains of the Arctic System Reanalysis (ASR). The outer domain has a resolution of 90 km and inner domain has a resolution of 30 km. Colors refer to the terrain height*.

4.4 Quality Control

Ground measurement of solar radiation is generally more prone to recording errors than other meteorological variables (96). For long time series assessment of estimating datasets, the quality of the ground measurement is very important. A close examination of the ground-measured solar radiation reveals that there are errors for extended periods of time (97). Younes, Claywell (97) identified two major types of errors in the ground measurements from pyranometers. The first type of error is called the *uncertainty of equipment* error, which is introduced because of the construction and calibration of the equipment. The second type of error is the operational error, which is related to the maintenance of the sensor. Because of the existence of such errors and their effects on the validation or feasibility studies, it is crucial to perform quality-control (QC) procedures on the solar radiation data (98). The ground-measured data used in this thesis is quality controlled by the respective organizations. In case of SMHI, Baseline Surface Radiation Network (BSRN) routines by Long and Dutton (67) are used for quality assurance. Missing or erroneous data is corrected by using meteorological variables described by Davies and McKay (99). The SMHI network was upgraded in 2006-2007 and the average correlation ratio between old and new measurements was found to be 0.997. More

*National Center for Atmospheric Research Staff (Eds). Last modified 09 Nov 2017. "The Climate Data Guide: Arctic System Reanalysis (ASR)." Retrieved from https://climatedataguide.ucar.edu/climate-data/arctic-system-reanalysis-asr.

detail on the upgrade is given by Carlund (65). SMHI provides data with quality flags and before using the data, these quality flags can be analyzed. In the case of NIBIO, the ground-measured data is quality controlled and the equipment is regularly maintained on a daily or weekly basis (66).

Although the data used here is quality controlled, Urraca, Gracia-Amillo (68) observed that operational and equipment errors exist especially in NIBIO stations. The first check performed in this regard is to look at the percentage of missing data. In Paper I and II, any year having more than 10% of missing data was discarded, however extra quality checks were not performed. In Paper III and IV, years having more than 5% of missing data were discarded. Moreover, the QC procedures described in the following sub-sections were performed in Papers III and IV.

4.4.1 BSRN Global Network recommended Quality Control test V2

The Baseline Surface radiation Network (BSRN) and its central archive – the World Radiation Monitoring Center (WRMC) provides the best possible quality controlled data for long- and short-wave surface solar radiation. To assure the quality, data received by WRMC/BSRN from ground-measuring stations runs through an inspection that includes the BSRN recommended quality checks V2.0 (67). The quality of the data is then represented in the form of flags (100). For global shortwave radiation, two tests are applied that check the *physically possible limits* and the *extremely rare limits*. The physically possible limits are shown in Equation 9 and the extremely rare limits are shown in Equation 10.

$$Min: \ -4\ Wm^{-2}$$

$$Max: S_a \times 1.5 \times \mu_o^{1.2} + 100\ Wm^{-2} \tag{9}$$

$$Min: \ -2\ Wm^{-2}$$

$$Max: S_a \times 1.2 \times \mu_o^{1.2} + 50\ Wm^{-2} \tag{10}$$

$$\mu_o = \cos(SZA) \tag{11}$$

$$S_a = {S_o}/{AU^2} \tag{12}$$

Where, SZA is the solar zenith angle, S_o is the solar constant at mean Earth-Sun distance and AU is the Earth-Sun distance in Astronomical units. After performing these quality control

tests, years having more than 1% of the flags were discarded from the analyses. The BSRN and similar tests are designed to detect only large deviations in ground-measured records; however, small errors introduced by shading, soiling, frost, snow or calibration of the equipment are not detected by these procedures (68).

4.4.2 Quality Control with Reanalysis and Satellite-based Products

As described in the previous section, general quality control (QC) procedures that principally test the range, model comparison, and graphical analysis are not effective in detecting small but persistent errors. Keeping this in view a more sophisticated QC procedure by Urraca, Gracia-Amillo (68) is presented here. This semi-automatic procedure is based on the statistical analysis of ground-measured solar radiation and radiation from reanalyses or satellite products. These products generally have larger errors than ground-measured data but operation and equipment errors are not as common in these as in ground-measured data. In the first step of this QC procedure, a confidence interval is constructed by calculating daily deviations (∂_t) of the products as shown in Equation 13.

$$\partial_t = Y_t - O_{t,} \tag{13}$$

where, Y_t are the estimations, O_t are the observed values and t is the temporal resolution. The confidence interval is then calculated for monthly values (temporal averaging) and for groups of stations with similar characteristic (spatial averaging). The averaging for the time and space is performed in two steps to increase the robustness of the confidence intervals. First, the bias with respect to median of daily deviations is calculated for each months and location as shown by Equation 14.

$$\widehat{Bias} = median\,(\partial) \tag{14}$$

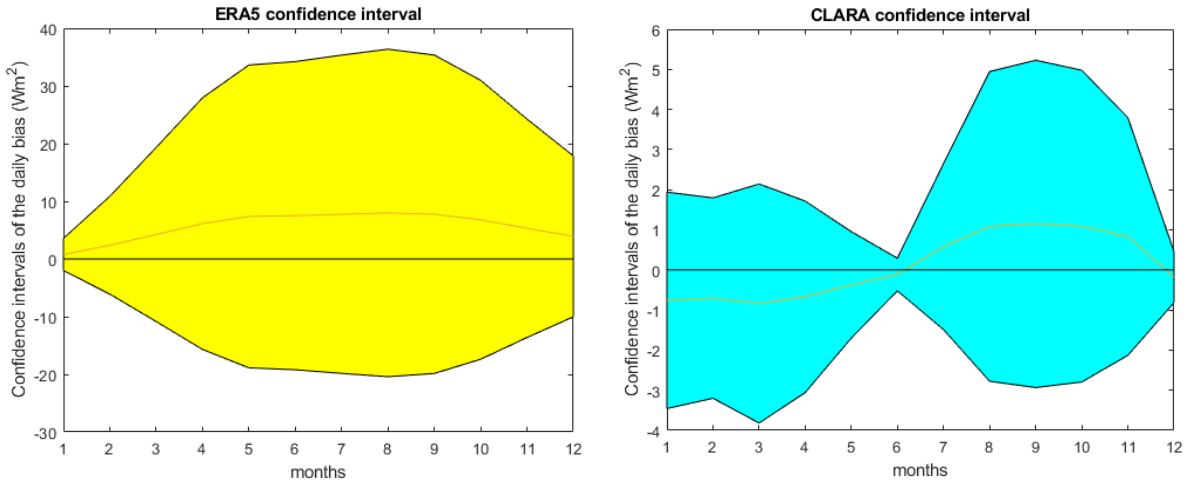
The Bias obtained from Equation 14 is again averaged on months of the year and stations within the same spatial group, resulting in a unique set of twelve values per group per product. To include the measure of dispersion, mean absolute deviation (MAD) is calculated by the following equation.

$$MAD = 1.4286 \times median\left(\left|\widehat{Bias}\right|\right) \tag{15}$$

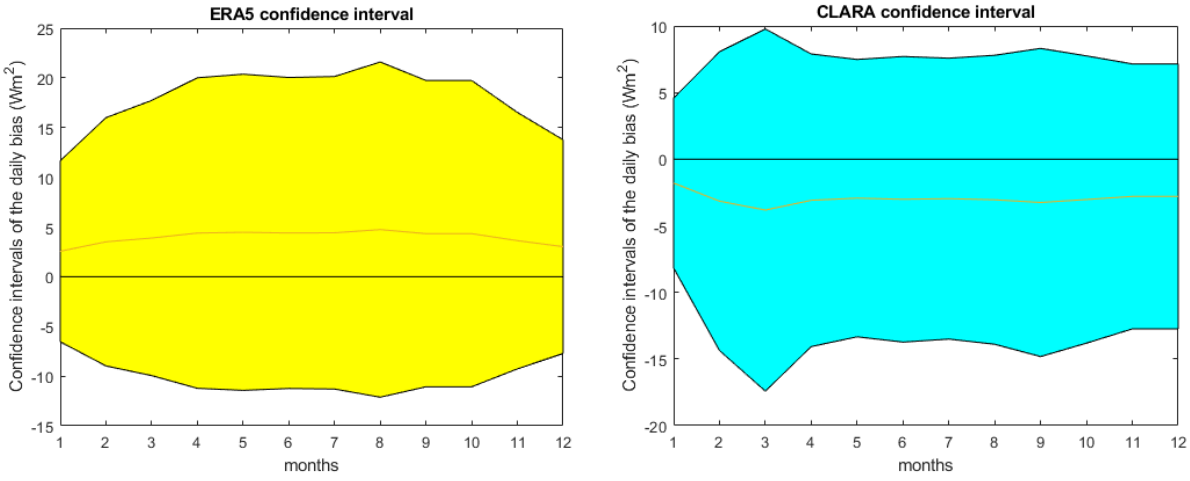
The *MAD* includes a constant scale factor of 1.4286 to ensure the consistency of estimates for different sample sizes. Finally, the confidence interval (*CI*) is calculated by the following equation.

$$CI = \widehat{Bias} \pm n \times MAD \quad (16)$$

Where *n* is a coefficient that weighs the *MAD* in order to adjust the level of QC procedure. The value of *n* is set to respectively, 2.4 or 0.4 for flagging operational errors and equipment errors. Figure 4.4 illustrates the confidence intervals developed in Papers III and IV.



(a) Above 65ºN locations in Norway



(b) Below 65ºN locations in Norway

Figure 4.4: The confidence intervals constructed for the quality control from ERA5 and CLARA. The locations in Norway were divided into two categories based on the latitude. Locations above 65ºN were placed in one group (a), and locations below 65ºN were placed in the second group (b). Based on the quality control procedures presented by Urraca, Gracia-Amillo (68) and adapted from Papers III and IV.

After constructing the confidence intervals, a window width parameter (*w*) is defined. As with the value of *n*, the window width can be set to either 20 for operational errors or 90 for equipment errors. The window starts increasing with a step of five days (fast moving filter),

and for each group of days it flags the data which is above or below the confidence interval. Days with an absolute relative deviation of 5% or absolute deviation of 5 Wm^{-2} are not accounted in flagging. The products that are more accurate provide a stricter confidence interval as can be seen in Figure 4.4. In the case here, as CLARA is a more accurate dataset, it has a much narrower confidence interval when compared with ERA5. Hence, CLARA will flag more data points than ERA5 because of the higher accuracy.

In the final step of this QC procedure, two graphical plots are generated for visual analysis. The first graph is generated for the daily deviations between the product and the ground data and a second graph is generated for comparing the hourly irradiance of ground measurements and product. For information on these graphs, refer to Urraca, Gracia-Amillo (68). Both of these graphical plots are examined visually to detect any false alarms. As the graphical comparison is performed for hourly averaged values, it is convenient to include at least one product that has hourly resolution (68). Initially, the locations Pasvik, Mære, Ullensvang, and Njøs were included in Paper III but after performing this QC test, large numbers of errors were found. These locations were discarded from this thesis and from Papers III and IV.

4.5 Random Forest Classification and Regression

Recently, there has been a growing interest in ensemble learning techniques. Ensemble methods are based on generating many classifiers and the results of these are aggregated which increases the learning ability for the entire inputs and target (101). Random forest regression (RFR) is a regression tree method, which has become very popular in recent years due to its strong performance, ease of implementation and low computational cost. It is an ensemble learning technique developed by Leo Breiman (102), which is based on the construction of a multitude of decision trees, where branches of the trees represent a particular path for the input data and leaves represent the output values. In RFR, a particular tree is grown in accordance with the realization of a random vector. The final prediction is based on aggregation over the ensemble of trees, referred to as the forest (103). On each of the trees, branches or nodes are made which are based on comparing a randomly selected feature to a random threshold. The randomness introduced in both variable selection and threshold determination has been shown to results in attractive properties such as a controlled variance, resistance to overtraining, and robustness to outliers as well as irrelevant variables. Moreover, RFR inherently provides estimates of generalization error and measures of variable importance (104, 105). The process of dividing the input data over branches are repeated until one or a pre-set number of data points are contained in each branch. This final node of the tree is referred to as a leaf, and it represents the

final-outcome of that particular regression in the whole model. The structure of the forest and hence the RFR behavior can be controlled by three main parameters, namely the number of trees (with a default value of 500), the number of variables considered in each node (generally set to m=P/3 following common practice in RFR), and the final number of data points that can make a leaf (our default value is 1). Having very low number of leaves in the model can cause overfitting, which can be overcome by pruning, *i.e.* limiting the number of data points in each leaf. With an increase in the number of trees, the computation load increases. An initial increase in the accuracy of the regression will also be observed, before reaching a saturation point (106), after which improvements are limited by a strong correlation between the trees (102). The RFR is used in Paper IV to construct a multi variate regression data set based on CLARA-A2 and ERA5 datasets.

4.6 Statistical Evaluation of Estimations

In order to evaluate the performance of the datasets, some common statistical measures were used. The most widely used measure is the Root Mean Squared Deviation (RMSD), which is given by Equation 17.

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(GHI_{estimated,i} - GHI_{observed,i}\right)} \quad (17)$$

Where, $GHI_{estimated,i}$ is the estimated global horizontal irradiance, $GHI_{observed,i}$ is the ground-measured global horizontal irradiance and N is the number of data points in time. As an additional measure, the MBD (Mean Bias Deviation) or bias was also used in the evaluation as shown in Equation 18. MBD gives an insight in the general trends of under or over estimations.

$$MBD = \frac{1}{N}\sum_{i=1}^{N}\left(GHI_{estimated,i} - GHI_{observed,i}\right) \quad (18)$$

Mean absolute bias deviation (MABD) was also used for the evaluations of datasets. Because of the absolute values used in this measure, the negative and positive errors do not cancel out each other as in MBD. This is a good measure to compare different models, as the one with smaller MABD will be the more reliable for estimations.

$$MABD = \frac{1}{N}\sum_{i=1}^{N}\left|GHI_{estimated,i} - GHI_{observed,i}\right| \quad (19)$$

The standard deviation of the error (STD) is used to evaluate the data set presented in Paper 4. The sample STD is computed as

$$STD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(\left(GHI_{estimated,i} - GHI_{observed,i}\right) - \left(\overline{GHI_{estimated}} - \overline{GHI_{observed}}\right)\right)^2}. \quad (20)$$

In addition, a bias-variance decomposition was used to obtain the optimal configuration of the random forest regression model used in Paper 4, with respect to the number of trees and the number of leaves. Moreover, R^2 and scatter plots were used to indicate the spread and overall correlation of the datasets with ground measurements.

4.7 Data extraction

The data extraction from the gridded datasets was performed in two ways. For high-resolution datasets like SARAH and ASR, the nearest grid point to the coordinates of the location was selected for data extraction. However, for coarse resolution datasets like ERA5 and CLARA, inverse distance weighting (IDW) interpolation was used. The IDW interpolation is given by the following equation.

$$\hat{V} = \frac{\sum_{i=1}^{n}\frac{1}{d_i}V_i}{\sum_{i=1}^{n}\frac{1}{d_i}} \quad (21)$$

Where, V_i are the known values, d_i are the distance from the data point and estimated point, and $\hat{V}$ is the value to be estimated. The four nearest surrounding grid points to the location were selected from ERA5 and CLARA as inputs to the IDW interpolation. Missing values exist in the CLARA dataset and if two or more of the surrounding four grid points were not available; the interpolation was replaced by a missing value.

4.7.1 Gap filling procedure

Gaps are often available in the ground-measurement and estimated surface solar radiation databases. Gaps in the ground measurement may occur due to power loss, misalignment, failure of instrument, insufficient cleaning or other reasons (107). In the satellite databases used here, the gaps in the data exist generally because of low number of observations and snow covers. In most of the analysis made here, gap-filling procedures were not used except in energy calculations in Paper III. In Paper III, nearest-neighbor interpolation was used to fill the gaps in SARAH, CLARA and ground-measured data. In addition, linear interpolation was used in filling gaps in Paper I.

5. Previous research and current knowledge gaps

In this section, an overview of the previous research done on estimating surface solar radiation is presented. In the last part of this section, the knowledge gaps are discussed from the perspective of estimating solar radiation in high latitude locations.

5.1 Previous research

Most of the research on remotely estimating solar radiation has been performed for mid- and low-latitude locations by using geostationary satellites. One of the earliest validation of these estimations was carried out by Hollmann, Mueller (108). In this research, the authors used the data from AVHRR sensor on-board polar orbiting satellites and showed that the average mean biases were small and were within the targeted accuracy of 10 Wm^{-2} on monthly mean time scales. In Posselt, Mueller (109), authors used geostationary Meteosat second generation satellites and evaluated the estimated radiation at 10 locations from the BSRN network. The highest latitude location analyzed in this study was Lerwick (UK) and this location had the highest mean absolute deviation (MABD). In a subsequent study (110), particular improvements were found at Lerwick because of the advancements in retrieval methods. The bias and MABD were remarkably low with 1.27 Wm^{-2} and 5.46 Wm^{-2}, respectively. In total, about 94% of the monthly mean values showed an accuracy of 10 Wm^{-2} or better. In Bojanowski, Vrieling (111), authors showed that the solar radiation estimation from Meteosat first and second generation satellites had a similar accuracy, however the authors suggest that ERA-Interim can be used as an effective backdrop to satellite products. In Sanchez-Lorenzo, Enriquez-Alonso (112), authors studied the trends in surface solar radiation over Europe from CM-SAF geostationary satellite products but high altitude locations were excluded from the study, because such locations are known to have problems in deriving surface solar radiation as shown in some other previous studies (113, 114). Similarly in Cristóbal and Anderson (115), the authors used Meteosat second generation satellites to estimate solar radiation over the northeastern Iberian Peninsula. In this study, it was observed that the errors were small in flat areas while an increase in errors was observed in mountainous regions. Another such study outlined the difficulties of satellites in estimating solar radiation in mountainous regions (116). In this study, three different algorithms were used to estimate surface incoming solar radiation in Belgium. Although, all the algorithms underestimated solar radiation when compared to ground measurements, the authors of this study expected the sensitivity to increase in regions with strong influence of mesoscale meteorology such as coastlines and highlands as compared

to mid latitude regions with a rather flat orography. The reason for this shortcoming is explained in Amillo, Huld (117), which showed that the accuracy of estimating effective cloud albedo (CAL) decreases towards the edge of the field of view of satellite, mainly because of very shallow angles. This slant-viewing angle introduces biases that tend to be larger near the edge of the satellite images, which start affecting the accuracy around ±65º latitude.

In Alexandri, Georgoulias (118), authors compared CM-SAF SARAH dataset with CERES (Cloud and the Earth's Radiant Energy System), GEWEX (Global Energy and Water Cycle Experiment), ISCCP (International Satellite Cloud Climatology Project) and ERA-Interim for Eastern Mediterranean. Overall, SARAH performed better than other datasets. Similarly in Urraca, Martinez-de-Pison (119), authors analyzed global horizontal irradiance from SARAH, ERA-Interim, interpolated ground-measurements (Ordinary kriging) and a statistical model called XGBOOST. In this study, 38 ground stations in central Spain were evaluated and it was found that SARAH provides better solar radiation estimates with low variability. Both of these studies showed that satellite products underestimate solar radiation. In another study it was shown that intermediate-sky conditions are overestimated while these overestimations increase further in overcast conditions, however areas affected by snow may have larger uncertainties (117). In some studies, around 5-10 Wm^{-2} of mean absolute deviations for monthly means was observed in geostationary satellite databases (77, 109, 110, 114, 120, 121). Most of the studies reported satellite methods to underestimate incoming solar radiation, besides some studies like Žák, Mikšovský (120) and Hakuba, Folini (122) that reported overestimation.

One of the more relevant studies to this thesis was done by Riihelä, Carlund (84). In this study, authors validated the first editions of SARAH and CLARA datasets over multiple locations in Sweden and Finland, spanning from 55º to 70ºN. Both datasets were found to have monthly mean accuracy better than 10 Wm^{-2} and a daily mean accuracy of 15 Wm^{-2}. SARAH was only able to provide coverage in southern Nordic regions because of its limited coverage. However, unlike CLARA, SARAH provide coverage on snow covered surfaces, although the 2nd edition of CLARA now provides more coverage on snow covers (30, 31). SARAH error characteristics were seen to have latitude dependence and errors increase with increasing latitude.

Another very recent and relevant study was done by Urraca, Gracia-Amillo (30), in which authors made an extensive evaluation of CM-SAF products including SARAH-2 and CLARA-A2 datasets. In this study, 313 ground stations were evaluated from several European countries, which included 29 stations from Norway. Satellite datasets underestimated at high latitudes while a slight overestimation was observed in southern regions. CLARA showed very good

temporal stability while keeping a small constant underestimation in majority of locations, however, the MABD in CLARA was larger than SARAH by 1-2 Wm^{-2}. ERA-Interim was found to have a constant positive overestimation and absolute errors almost double that of the satellite datasets. In this study, although CLARA underestimated solar radiation, a significant decrease in the bias is found when compared to the first edition of CLARA dataset (81). Locations with seasonal snow covers, which are abundant at high latitude locations, were observed to have large underestimation. In a similar way, SARAH was seen to be underestimating as well but the underestimation was again larger for regions with snow covers. This is because the satellite algorithms only use the visible channel to detect the presence of clouds, hence these cannot differentiate if a bright pixel corresponds to a cloud or to a surface covered with snow. Moreover, satellite models fail on mountainous regions because the spatial and temporal resolutions are not high enough to account for the sharp terrain and changing weather conditions (30).

On the contrary, reanalysis overestimate incoming solar radiation as reported in multiple studies (30, 111, 119, 123). Although, not as many studies have been performed on the evaluation of reanalyses for incoming solar radiation as there are on satellite estimations, some studies like Urraca, Huld (28), Bojanowski, Vrieling (111) suggest that reanalysis have been improving and these can be used where the satellite data is missing or inaccurate.

5.2 Thesis work in relation to knowledge gaps

The previous section highlighted a number of interesting topics that were chosen for further research in this thesis. The knowledge gaps associated with these research areas are summarized and linked to the appended papers in the following:

- The number of meteorological stations recording shortwave incoming solar radiation is very low in Northern Norway. Even though there are many meteorological stations recording other atmospheric variables like temperature, precipitation, and humidity, the number of station recording solar radiation remains low. The model proposed in Paper I can be used to construct estimated solar radiation by using temperature and humidity at these stations.
- There are very few studies carried out on evaluating solar radiation datasets from polar orbiting satellites. Paper II and III provide an evaluative analysis for polar orbiting CLARA dataset for high latitude locations.

- Arctic system reanalysis (ASR), which is a polar optimized dynamic downscaling of ERA-Interim, was not evaluated any further for solar radiation in high latitude regions. An assessment was provided in Paper III on ASR version 2. It was found that this dataset provides very large uncertainties in estimating solar radiation.
- Because of the low coverage provided by geostationary satellites, they do not provide coverage in northern Norway. Moreover, the errors in geostationary datasets increase with increasing latitudes. Databases from polar orbiting satellites (CLARA-A1) can be used at high latitudes but because of snow covers, they have a large number of missing data as shown in Paper II.
- New datasets based on polar orbiting satellites (CLARA-A2) provide less missing values but these improvements are mainly on high latitudes and snow cover periods. When analyzed, these new data points were seen to have large errors as shown in Paper II.
- ERA5, a recently published reanalysis, is evaluated in Paper III. The results show that ERA5 provides reasonable errors and can be used as a supporting dataset when satellite datasets do not provide coverage, have missing values or large uncertainties.
- Reanalyses are reported to overestimate solar radiation while satellite databases underestimate solar radiation. A new dataset is presented here which is constructed by using a Random forest regression on reanalysis and satellite dataset. This model improves the solar radiation estimations in a number of ways. In the proposed model, there are no missing values, and the accuracy is better than both the reanalysis and satellite datasets.

6. Results

This chapter summarizes the results from the appended papers in two sections. In the first section, available resources of solar radiation estimation are analyzed and discussed. In the second section, the results from the evaluations of the datasets together with the use of a regression algorithm are used to create a novel and improved solar radiation dataset.

6.1 Evaluation of available datasets of surface solar radiation at high latitudes

This section summarizes the results from Papers I, II and III. In Section 6.1.1, a model based on the difference between maximum and minimum temperatures and relative humidity is presented from Paper I. In Section 6.1.2, the results from Paper II are presented which are based on a comparative analysis of CLARA-A1 and CLARA-A2. In Section 6.1.3, an analysis is presented on the estimation accuracies of CLARA-A2, SARAH-2, ERA5, and ASR from Paper III.

6.1.1 A model to estimate surface solar radiation by using temperature and humidity

This section provides an overview of the model developed in Paper I. The proposed model is based on the Hargreaves, Samani (76), in which authors have used the maximum temperature difference and extraterrestrial radiation in a day to estimate surface solar radiation, and eventually the evapotranspiration. The model presented by the same authors is shown in Equation 22.

$$R_s = K_{RS} R_a TR^{0.50} \tag{22}$$

Where, K_{RS} is an empirical coefficient fitted to R_s/R_a versus TR data, TR is the diurnal temperature difference between the maximum recording and the minimum recording, R_a is the extraterrestrial radiation, and R_s is the surface solar radiation. The value of K_{RS} in Equation 21 can take two different values, one for interior, and one for coastal regions. A value of 0.162 is recommended for interior regions and a value of 0.19 is recommended for coastal regions. The extraterrestrial radiation is calculated by the following equation.

$$R_a = \frac{24}{\pi} R_{sc} \left(1 + 0.33 \times cos \frac{360 \times P}{365}\right) \times cos(\varphi) \times cos(\delta) \times sin(hs) + \frac{(2 \times \pi \times hs)}{360} \times sin(\varphi) \times sin(\delta) \tag{23}$$

Where, R_a is the extraterrestrial radiation, R_{sc} is the solar constant with a value of 1366 Wm^{-2}, P is the day number (ranging from 1 for the first day of the year and 365 for the last day of the year), φ is the latitude, δ is the declination angle and *hs* is the hour angles of sunrise and sunset.

The model developed in Paper I is based on the model shown in Equation 22, but in addition to the temperature difference, relative humidity was taken into account. The proposed model is shown in the following equation.

$$R_s = 0.04 \times R_a \times TR + K_{RS} \times R_a \times (RH)^{0.27} \quad (24)$$

Where RH is the relative humidity in Equation 24. The empirical constant K_{RS} in Equation 24 can take two values, like the model presented in Equation 22 (76). A K_{RS} value of 0.01 is suggested for inland regions and a value of 0.04 is suggested for coastal regions. Figure 6.1 depicts the model estimated, observed, and extraterrestrial radiation for Tromsø, Norway in 2014.

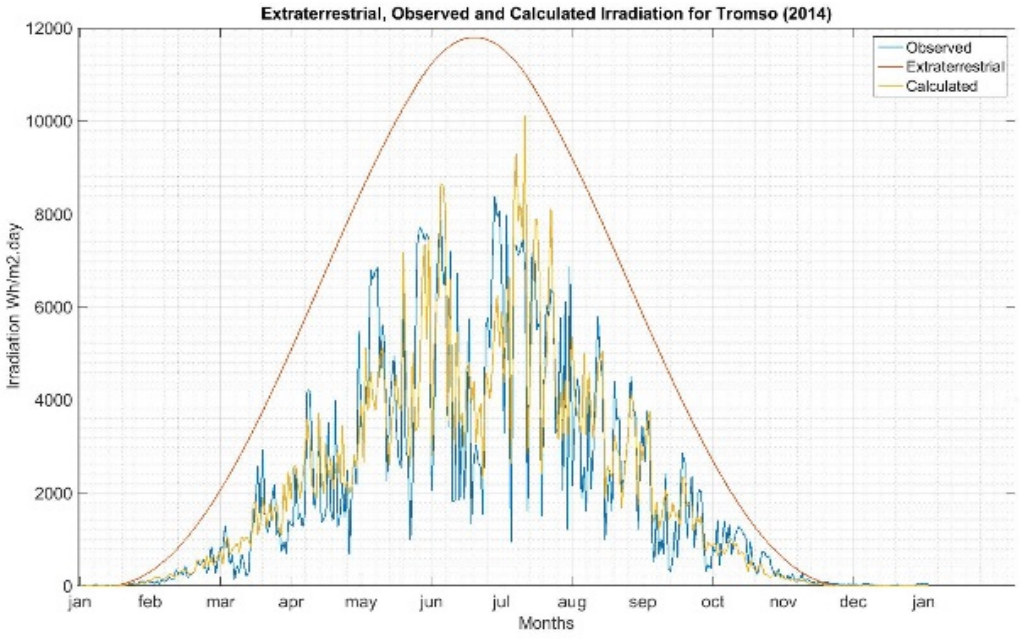


Figure 6.1: The model estimated GHI, observed GHI, and extraterrestrial radiation for Tromsø in 2014. The radiation is expressed in $Wh.m^{-2}$ (energy).

The model in Equation 24 was tested at eight locations in Norway. Compared to the original method proposed by Hargreaves and Samani, the daily average percentage error was improved by 0.2%, and yearly average percentage error was improved by 10.8%.

6.1.2 A comparison of CLARA datasets and an analysis of improvements in CLARA-A2

Most solar radiation datasets do not provide coverage above 65ºN (or below 65ºS) because majority of these datasets are based on geostationary satellites (an example of which is the SARAH dataset, discussed in the next section). For areas above 65ºN, the CLARA datasets, published and managed by CM-SAF, provide precise surface solar radiation estimations. At the time of writing, CM-SAF has published two editions of CLARA datasets. For further information on these datasets please refer to Section 4.3.2 or Karlsson, Anttila (81) and Karlsson, Riihelä (77).

In Paper II, a comparative analysis was presented for CLARA-A1 and CLARA-A2 datasets with an emphasis on the improvements of CLARA-A2. The study was performed for eight locations in Norway and seven locations in Sweden for 14 years between 1995 and 2009. The ground data for the analysis was acquired from NIBIO and SMHI, but unlike in Papers III and IV, quality control procedures were not applied except discarding years with more than 10% of missing values.

In this analysis, it was observed that the new dataset (CLARA-A2) had less missing data points; however, the errors and biases were found to be reduced in the previously existing data points when compared to CLARA-A1. Figure 6.2 shows the Hovmöller plots for CLARA-A1 and A2 datasets, which highlights the quantity of missing data points in each dataset.

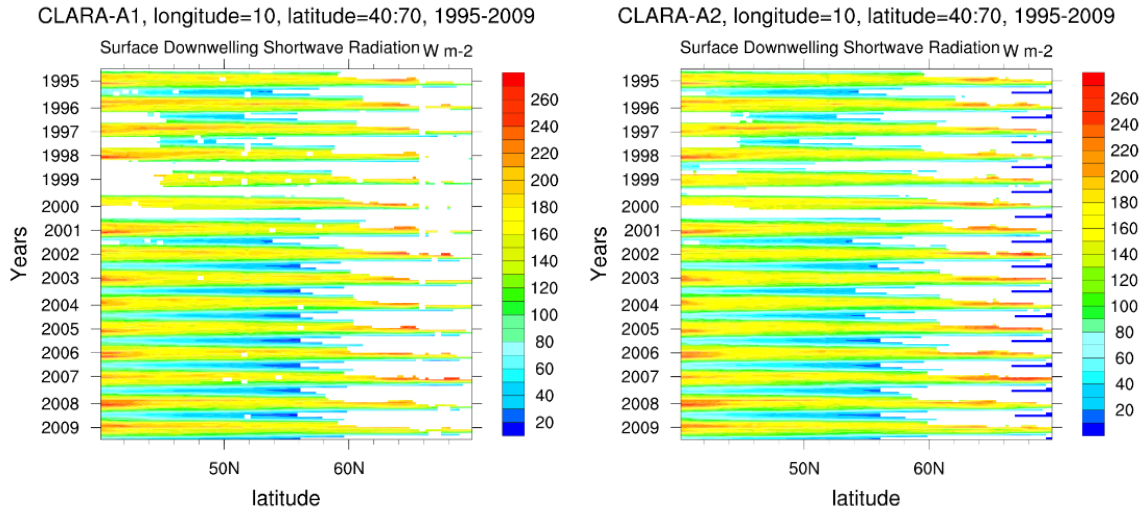
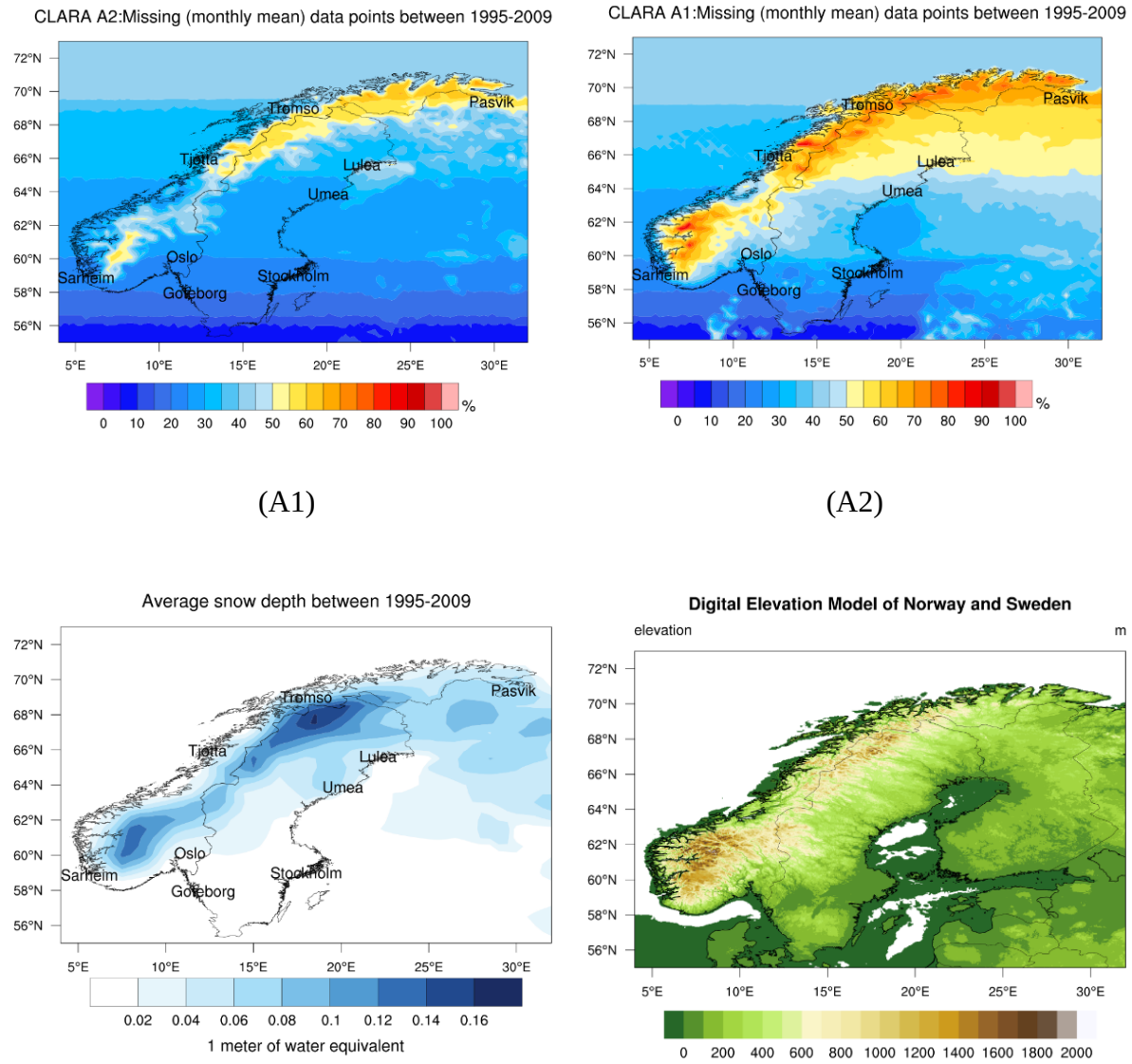


Figure 6.2: Hovmöller plots for CLARA-A1 and A2 datasets for 1995 to 2009. These plots are centered at 10º longitude and span from 40ºN to 70ºN latitude.

As seen from Figure 6.2, both the datasets have increasing number of missing values with increasing latitudes (latitudes increase from left to right in Figure 6.2). CLARA-A2 had less number of missing data points than the previous edition CLARA-A1. However, as can be seen from Figure 6.3, the improvement in the data availability is mostly on the high latitude areas that have more snow depth than the low latitude areas.



(A1) (A2)

(B1) (B2)

Figure 6.3: Missing data in CLARA-A1 and A2 datasets is illustrated. From (A1) and (A2) it can be seen that the number of missing data is reduced in CLARA-A2, however the decrease in missing data is mostly on high latitude locations with high snow depths, as shown in (B1) and (B2). From Paper II.

The increase in the availability in CLARA-A2 was mostly in snow-covered regions. As explained in Section 4.3.2, satellite estimation methods particularly those used in CLARA datasets have difficulties in differentiating between clouds and snow-covered surfaces because IR channels are not used in the radiation estimation algorithm. These new data points had very large errors especially at the locations studied in Paper II. Norwegian locations had a 12% increase in the availability of data and Swedish locations had a 9.6% increase, and as can be

seen from Figure 6.3 (B1), Norwegian locations receive more snow than Swedish location. Figure 6.4 depicts the increase in availability in CLARA-A2 dataset in quarter-yearly monthly averages. In the period from February to April, coastal regions in Norway and central parts of Sweden had the most increase in data availability. While in the period from May to July, the inland and southwestern parts of Norway and northern parts of Sweden had the largest increase.

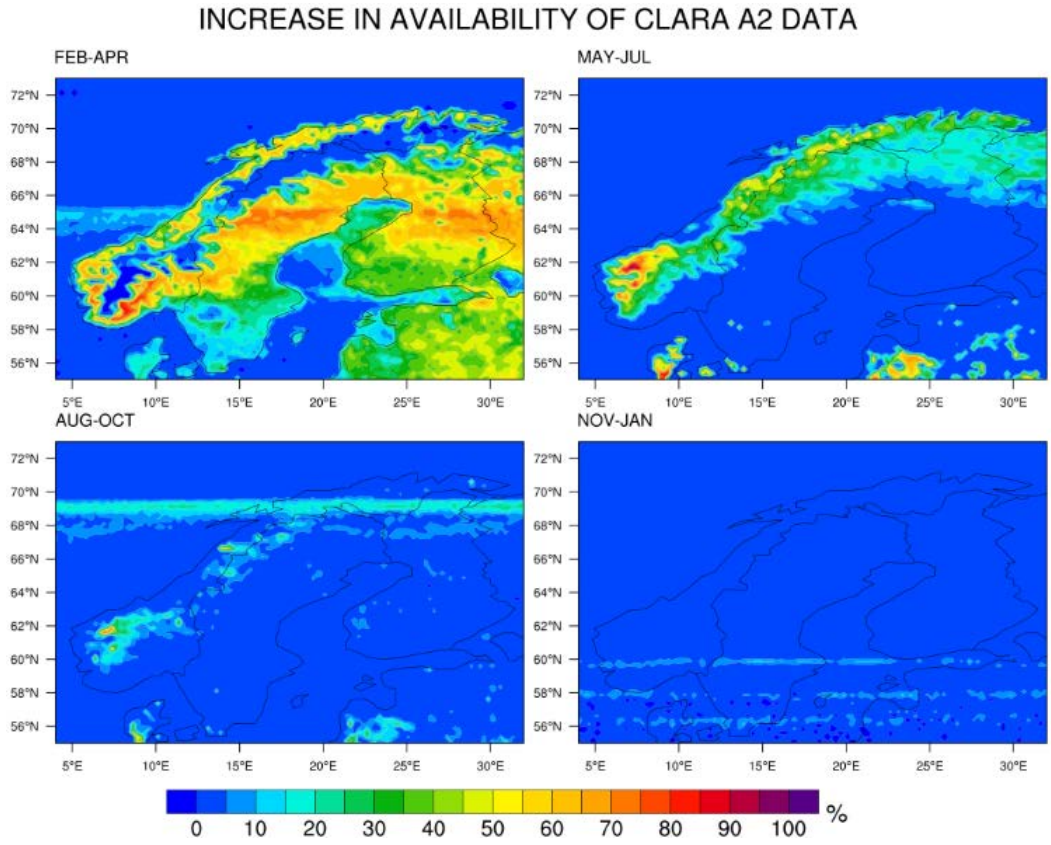


Figure 6.4: Percentage increase in the availability of CLARA-A2 dataset in each quarter. The highest increase is in the areas that have complex topography in addition to snow covers.

For Norwegian locations, the new data points had a mean absolute bias deviation (MABD) of 17.7 Wm^{-2} while for Swedish locations, an MABD of 15.2 Wm^{-2} was found. In comparison to the errors in new data points, other data points had an MABD of 8.3 Wm^{-2} for both Norwegian and Swedish locations. This showed that the new data points had large errors because these are primarily estimated on snow covers.

Overall, CLARA-A1 had an MABD of 8.0 Wm^{-2} and CLARA-A2 had an MABD of 8.9 Wm^{-2} for Norwegian location. For Swedish locations, CLARA-A1 had an MABD of 8.1 Wm^{-2} and CLARA-A2 had an MABD of 8.7 Wm^{-2}. However, for all location including sites from Norway and Sweden, CLARA-A1 had an MABD of 8.0 Wm^{-2} and CLARA-A2 had an MABD of 8.8

Wm^{-2}. The MABD was observed to be larger in the new edition of CLARA because as previously explained, this dataset had less number of missing values, and the new values were mostly on the snow-covered regions, which increased the overall errors in the dataset.

In Paper II an energy analysis for CLARA-A1 and A2 was performed. The energy is expressed in kWh on a meter square in a year and it was calculated by integrating the daily average values. In this particular analysis, gap filling was not applied. Evidently, as CLARA-A2 had less missing values than CLARA-A1, it was found that CLARA-A2 estimated yearly energy values more accurately than CLARA-A1. The conclusion drawn from this study was that CLARA-A2 brings improvements but at the cost of high errors on the new data points which were previously not available in CLARA-A1.

6.1.3 Investigating solar radiation datasets for high latitude locations – A comparative analysis of CLARA-A2, SARAH-2, ERA5 and ASRv2

In Paper III, CLARA-A2, SARAH-2, ERA5, and ASRv2 datasets were analyzed for their accuracy at 31 locations in Norway. The coordinates and land type of locations included in the study can be found in the Appendix, Table A. In addition to accounting for the accuracy, this study also gives a comparative analysis for the surface solar radiation datasets for high latitude locations. In Paper III, three quality-control procedures were applied as described in Section 4.4. In the first control, years having more than 5% of missing data were removed from the analysis. A second quality control was applied by using BSRN Global Network recommended Quality Control test, V2.0 (67) as explained in Section 4.4.1. A final quality control procedure is applied based on Urraca, Gracia-Amillo (68), which is explained in Section 4.4.2. For a list of year not included in the study, refer to the Appendix, Table B. Table 6.1 shows the properties of the datasets used in this study.

Table 6.1: Description of the datasets used in this study. The period analyzed, spatial, and temporal resolutions are shown for each dataset.

Datasets	Method	Years analyzed	Spatial resolution	Highest temporal resolution	Spatial limits
CLARA	Polar-orbiting Satellite	2000-2015	0.25ºx0.25º	Daily	Global
SARAH	Geostationary Satellite	2000-2015	0.05ºx0.05º	30 min	Limited to ±65º latitude and ±65º longitude
ERA5	Reanalysis (Global)	2000-2015	0.281ºx0.281º	Hourly	Global
ASRv2	Reanalysis (Regional renalaysis downscaled from ERA-Interim)	2000-2012	0.136ºx0.136º	3 Hours	180W - 180E longitude 24.643N - 90N latitude

The datasets were assessed based on RMSD, MABD, and MBD for daily, monthly, and yearly averages of GHI. In addition, a yearly energy analysis was performed. To assess the accuracy for different geographical regions, the locations were divided into four categories, as explained in Section 4.2. Moreover, a sky stratification analysis was performed to assess the performance of these datasets in different sky conditions. In the end, ERA5 was analyzed in-depth for cloud placement by investigating the total column of water content and agreement on sky classification by comparing it to ground-measured data and CLARA-A2 dataset.

Table 6.2: Error metrics expressed in Wm^{-2}, for the datasets analyzed in Paper II. Numbers without parentheses are monthly averaged errors while those in parentheses are daily averaged errors. Numbers are averaged over all stations. Error metrics for different geographical groups are also shown.

	RMSD (Wm^{-2})				**MABD (Wm^{-2})**				**MBD (Wm^{-2})**			
	CLARA	**SARAH**	**ERA5**	**ASR**	**CLARA**	**SARAH**	**ERA5**	**ASR**	**CLARA**	**SARAH**	**ERA5**	**ASR**
All Sites	9.5 (18.2)	8.7 (17.9)	9.9 (26.4)	21.7 (42.6)	6.3 (12.6)	5.8 (11.6)	6.4 (16.7)	14.5 (27.1)	-3.1 (-1.7)	-3.6 (-2.6)	2.1 (4)	13.1 (16.9)
Above 65ºN	10.1 (16.0)	-	10.9 (26.3)	20.3 (39.4)	5.4 (9.7)	-	6.1 (14.5)	11.1 (21.5)	-3.4 (-2.9)	-	3.7 (5.6)	8.0 (11.0)
Below 65ºN	9.3 (18.4)	8.7 (17.9)	9.8 (26.5)	21.9 (43.0)	6.4 (13.0)	5.8 (11.6)	6.4 (17.0)	15.0 (27.9)	-3.0 (-1.5)	-3.6 (-2.6)	1.9 (3.8)	13.8 (17.8)
Coastal	9.1 (16.9)	8.6 (17.1)	10.0 (26.4)	21.8 (41.9)	5.8 (11.6)	5.7 (11.2)	6.2 (16.3)	13.9 (25.6)	-2.8 (-1.4)	-3.5 (-2.3)	2.2 (4.2)	11.9 (15.7)
Inland	9.8 (19.1)	8.8 (18.1)	9.9 (26.4)	21.7 (43.1)	6.7 (13.4)	5.8 (11.9)	6.5 (17.1)	15.0 (28.3)	-3.3 (-1.2)	-3.7 (-2.8)	2.1 (4.0)	14.0 (18.0)

From Table 6.2, it can be seen that CLARA and SARAH are more accurate than ERA5 and ASR. ASR was observed to have very low accuracy when compared to other datasets, partly because it is a downscaling of ERA-Interim, which is a predecessor of ERA5. On location above 65ºN, CLARA had smallest errors among all datasets. On monthly averages, CLARA provided an MABD of 5.4 Wm^{-2}, whereas ERA5 had a MABD of 6.1 Wm^{-2}. ASR had a large MABD of 11.1 Wm^{-2}. SARAH being a dataset based on geostationary satellites does not provide coverage above 65ºN. However, at location below 65ºN, SARAH had the smallest MABD of 8.7 Wm^{-2}, followed by CLARA with an MABD of 9.3 Wm^{-2} and ERA5 with an MABD of 9.8 Wm^{-2}. ASR again had the largest MABD among the datasets with 15.0 Wm^{-2}. In coastal and inland locations, a very similar pattern was observed where SARAH performed better than other datasets. However, in inland regions, ERA5 had slightly smaller error than CLARA; because most of the inland regions of Norway receive more snow cover when compared to coastal regions (see Figure 6.3 (B1)). In agreement with many previous studies, this analysis found satellites databases to underestimate solar radiation and reanlyses to overestimate solar radiation.

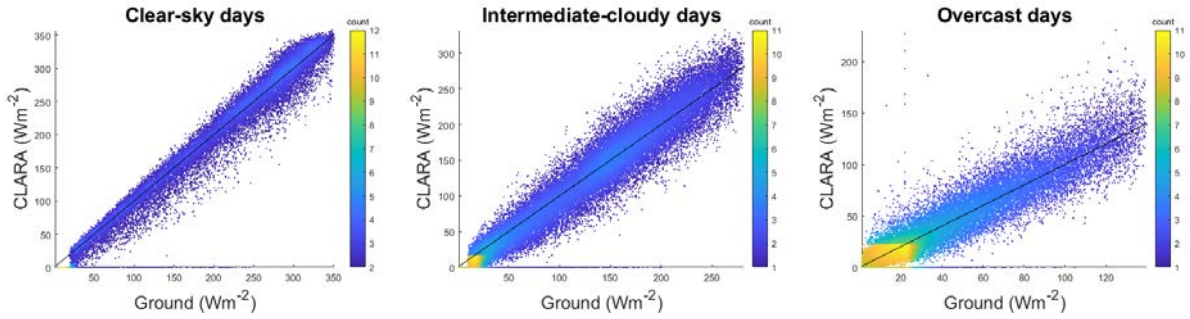
One of the main challenges of estimating surface solar radiation from any method is the accurate placement of clouds in time and space. However, even the most accurate and sophisticated methods fail to accurately estimate clouds in clear-sky and cloudy conditions. To assess the sky stratification accuracy of the datasets studied, a clear-sky index was used. The clear-sky index is defined as the ratio of clear-sky GHI to the GHI recorded on the ground, given by the following equation.

$$CSI = \frac{GHI_{ground}}{GHI_{clear-sky}} \tag{25}$$

Where, CSI is the clear-sky index, GHI_{ground} is the global horizontal irradiance observed on ground and $GHI_{clear\text{-}sky}$ is the global horizontal irradiance from a clear sky model. For sky classification of these datasets, the Bird clear-sky model was used (124). After calculating clear-sky indices, following Smith, Bright (125) and Widén, Shepero (126), values larger than 0.8 were considered indicating a clear-sky day, values of CSI between 0.4 and 0.8 were considered as intermediate-cloudy and values below 0.4 were considered as overcast. This type of categorization is quite arbitrary in the sense that the actual conditions can vary to some degree, *e.g.* CSI values larger than 0.8 are categorized as clear-sky but a small amount of clouds may be present in any of the days in this category. Similarly, values below 0.4 are categorized as

overcast conditions but some days may have intermediate clouds. The main aim of making such a grouping was to separate the days into different types to assess the model performances. This can be seen in Figures 6.5 to 6.8 that the days categorized as clear-sky have larger maximum solar irradiance while days categorized as overcast have much smaller maximum solar irradiance.

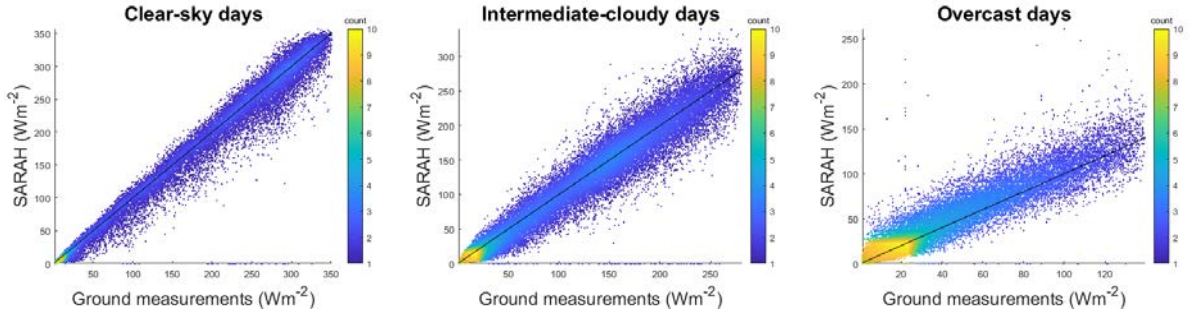
Figures 6.5 to 6.8 show the scatter plots of CLARA, SARAH, ERA5, and ASR datasets. These figures also list the RMSD, MABD, and MBD of these datasets in different sky categories. Overall, in the three categories, SARAH performed better than other datasets while ASR performed the worst. In clear-sky category, an underestimation was observed in SARAH, CLARA, and ERA5, while ASR overestimated radiation. Similarly, in the intermediate-cloudy category, both satellite databases underestimated, while reanalysis overestimated. Finally, in the overcast category, CLARA slightly underestimated solar radiation while other datasets overestimated. In conclusion, all the models were found to have discrepancies in presenting clouds in all types of sky conditions.



(a) (b) (c)

CLARA	RMSD (Wm^{-2})	MABD (Wm^{-2})	MBD (Wm^{-2})
Clear-sky	21.6	13.8	-4.1
Intermediate-cloudiness	22.2	16.0	-3.4
Overcast	13.8	8.7	-0.2

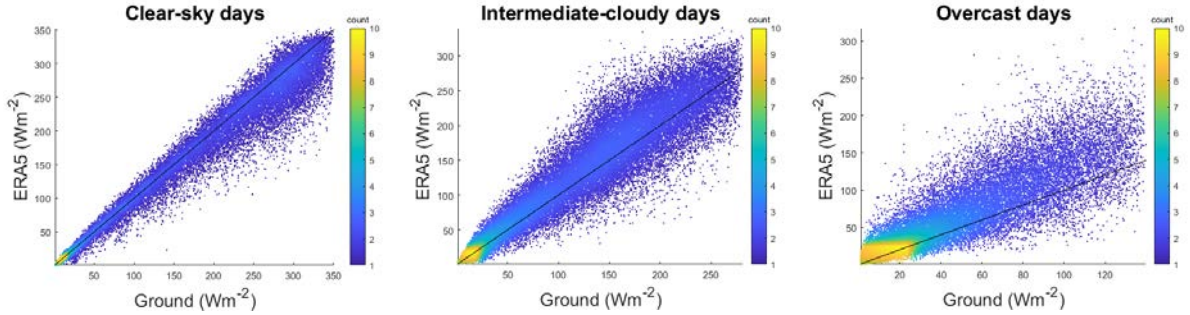
Figure 6.5: CLARA daily errors under clear-sky, intermediate-cloudiness, and overcast conditions. Scatter plots for different sky-conditions are shown. The colored legend bar shows the density of points in the scatter plot. From Paper III.



(a) (b) (c)

SARAH	RMSD (Wm^{-2})	MABD (Wm^{-2})	MBD (Wm^{-2})
Clear-sky	20.5	12.8	-5.6
Intermediate-cloudiness	20.2	13.5	-3.1
Overcast	13.3	8.7	4.4

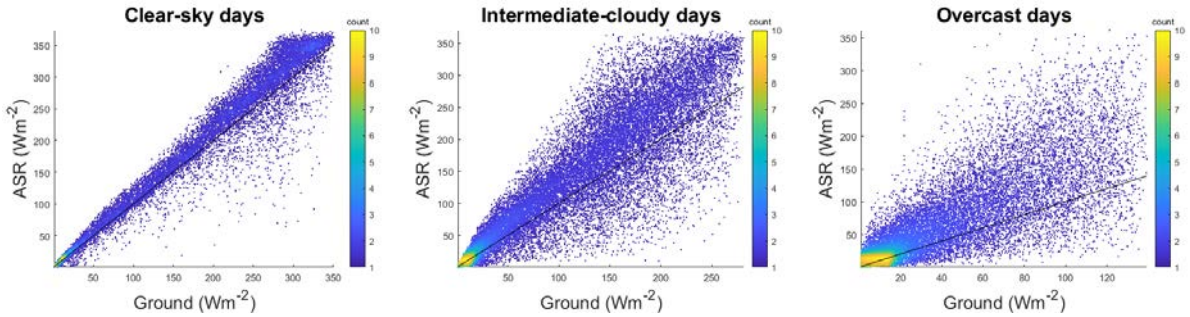
Figure 6.6: As in Figure 6.5 but for SARAH.



(a) (b) (c)

ERA5	RMSD (Wm^{-2})	MABD (Wm^{-2})	MBD (Wm^{-2})
Clear-sky	25.5	16.8	-10.0
Intermediate-cloudiness	28.4	19.8	8.7
Overcast	29.7	18.7	15.3

Figure 6.7: As in Figure 6.5 but for ERA5.



(a) (b) (c)

ASR	RMSD (Wm^{-2})	MABD (Wm^{-2})	MBD (Wm^{-2})
Clear-sky	29.2	21.1	11.6
Intermediate-cloudiness	51.3	37.2	23.3
Overcast	49.0	30.8	25.0

Figure 6.8: As in Figure 6.5 but for ASR.

Some shortcomings of satellite models in underestimating clear-sky and intermediate-cloudy conditions are explained here. Under clear-sky conditions, CLARA uses aerosol information

from Global Aerosol Data Set/Optical Properties of Aerosols and Clouds (GADS/OPAC) climatology and SARAH uses aerosol information from Monitoring Atmospheric Composition and Climate (MACC climatology). Both the datasets use integrated water-vapor information from ERA-Interim. Aerosol information from MACC climatology is observed to have higher accuracy than GADS/OPAC climatology (126). The maximum aerosol optical depth (AOD) is reduced in GADS/OPAC climatology for the CLARA dataset, but the results show that the climatology used in SARAH performs better than in CLARA even after the modifications. The negative biases observed in the clear-sky and intermediate-cloudy categories are possibly due to the aerosol climatology being too thick, which results in an underestimation of solar radiation. As reported in Mueller and Träger-Chatterjee (127) and Polo, Antonanzas-Torres (128), both MACC and GADS/OPAC climatologies cause an underestimation in surface solar radiation because of the apparent overestimation in AOD thickness. In addition to aerosol optical depth, vertically-integrated water vapor values taken from ERA-Interim are shown to be too large (129), which can further attenuate the surface solar radiation. In ERA5, the radiative transfer model RTTOV11 (Radiative Transfer for TOVS) has a tendency to underestimate reflectance of high cumulus cloud tops while the reflectance of lower water clouds is overestimated which can cause an underestimation in clear-sky conditions and overestimation in intermediate-cloudy and overcast conditions. In ASR, all the conditions are overestimated which shows that there is an underestimation in aerosol optical depth and cloudiness in the atmosphere.

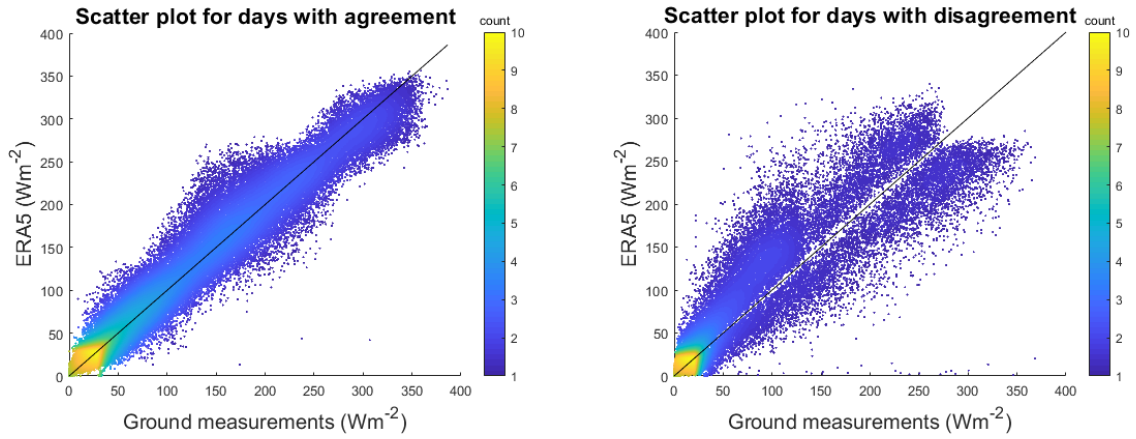
In the final analysis of this study, the cloud estimation accuracy of ERA5 was explored, as it is proposed as a complimenting alternative to satellite datasets. For all the locations, the RMSD of monthly values for ERA5 is similar to that of CLARA and SARAH, but the RMSD of daily values (in parentheses) was considerably larger in ERA5 when compared with the satellite databases. On even larger time scales (see Paper III), the difference decreased further. In this analysis, the total cloud water content (TCWC) and short wave solar radiation downward, clear-sky (SWSDC) from ERA5 were used here. Clear-sky indices for ground-measured data, ERA5, and CLARA-A2 were calculated by using SWSDC from ERA5 because the clear-sky values from ERA5 have the aerosol and water content information, which is used in calculating the surface solar radiation. This analysis was performed for days when the solar zenith angle is lower than 90º. Times when the solar zenith angle is higher than 90º was not considered in this analysis, as the intent here is to analyze solar radiation and TCWC, however, when the solar

radiation is not available, the TCWC is present. Including nighttime values in this analysis would have influenced these results.

Table 6.3: The number of days and mean TCWC from *in-situ* ground measurements, ERA5 and CLARA are shown in the table for different sky categories. The number of days and mean TCWC in each cloudiness category for ERA5 is shown separately for cases when ERA5 and ground measurements agree on classification and for cases when there is a disagreement. Years from 2000 to 2015 were used in this analysis over all locations included in the study.

	Ground data		**CLARA data**		**ERA5 data**		**ERA and ground agree**		**ERA and ground disagree**	
	No. of days	**Mean TCWC (Kg.m-2)**	**No. of days**	**Mean TCWC (Kg.m-2)**	**No. of days**	**Mean TCWC (Kg.m-2)**	**No. of days**	**Mean TCWC (Kg.m-2)**	**No. of days**	**Mean TCWC (Kg.m-2)**
Clear-sky	38265 (30.2%)	0.03	39516 (31.3%)	0.03	53211 (33.4%)	0.02	29500	0.02	8765	0.07
Intermediate-cloudiness	49207 (38.8%)	0.09	45244 (35.8%)	0.10	75268 (47.4%)	0.10	34700	0.10	14507	0.07
Overcast	39181 (30.9%)	0.22	41417 (32.8%)	0.22	30389 (19.1%)	0.29	20914	0.30	18004	0.12

In this analysis, it was found that ground measurement and CLARA classify almost the same percentage of days into each category, however, ERA5 was observed to classify a large number of days as intermediate-cloudy and a small number of days as overcast than *in-situ* observations, hence showing that it had a negative bias towards classifying a day as overcast. CLARA had very similar mean TCWC values as ground measurements but ERA5 slightly underestimated TCWC in the clear-sky category but largely overestimated it in overcast category, as shown in Table 6.3. Moreover, in ERA5 the mean TCWC was slightly underestimated in the clear-sky category but largely overestimated in overcast category. The agreement on sky conditions was also analyzed and it can be seen from Table 6.3 that the mean TCWC of days with agreement is the same as that of ERA5, but on the days of disagreement, there is an overestimation in mean TCWC in clear-sky days and an underestimation in overcast days. These results showed that on clear-sky days, ERA5 had more clouds than *in-situ* observations, which was seen by higher levels of TCWC, while on the overcast days there was a lower amount of clouds, which was seen by lower levels of TCWC. Figure 6.9 shows the scatter plot of ground measurements and ERA5 for both of these conditions, *i.e.* when there is an agreement on classification and when there is a disagreement. It can be seen that the spread is large when there is a disagreement. A correlation coefficient of 0.98 is found for agreement data points while a correlation coefficient of 0.90 is found for disagreement points.



(a) (b)

Figure 6.9: Scatter plots for the days when ERA5 and ground measurement agree on classification and when there is a disagreement. A correlation coefficient of 0.98 is found for agreement points and 0.90 for disagreement points.

The RMSD, MABD, and MBD were calculated for different sky conditions and when ERA5 and ground measurements agreed on sky conditions and for when there was a disagreement. This error analysis showed that the highest increase in errors was seen in clear-sky and overcast categories with MABD of 42.6 Wm^{-2} and 30.6 Wm^{-2}, respectively. The MBD was positive in clear-sky category and negative in intermediate-cloudiness and overcast categories, which further showed that there was less amount of clouds in the clears-sky category and more amount of clouds in intermediate-cloudiness and overcast categories. From a solar energy-harvesting point of view, the clear-sky days produce more energy than intermediate-cloudy or overcast days. It can be observed that ground-measurement and ERA5 predicts almost the same percentage of clear-sky days, which further shows that on daily averages, reanalyses may not predict clouds accurately but on longer time scales, the solar radiation estimation improves.

In conclusion, both CLARA and SARAH provided good estimates but both of these datasets had some shortcomings, including the spatial limits of SARAH and the low temporal frequency of CLARA. On the other hand, ERA5 provided advantages in the form of historical data series and global coverage. Based on these results, it was suggested that CLARA and SARAH provide better estimates for solar radiation, but ERA5 can be used to fill the missing data in these datasets.

6.2 A Random Forest regression based model

As presented in previous section, satellite based models are more accurate than reanalyses, however the accuracy of satellite models deteriorate with increasing latitude. Moreover, unlike reanalyses, satellite models have missing values and a negative bias. In Paper IV, a novel method was presented which is based on taking advantage of these over and underestimation of ERA5 and CLARA datasets. A regression-based method was used to construct a new datasets by using CLARA and ERA5. The new dataset provided more accurate estimations of surface solar radiation than the input datasets.

The regression model used in Paper IV is called Random Forest Regression (RFR), explained in Section 4.5. Initially in this study, Gaussian process regression was used to improve the solar radiation estimates, but experimenting with RFR provided better results. In this study, 31 locations from NIBIO solar radiation-measuring network were used (refer to the Appendix, Table A for information on the locations and Table B for information on rejected years). In addition, five stations from SMHI solar radiation measuring network from Sweden were used to evaluate the performance of the proposed dataset (Appendix, Table C). To train the model, 20% of the data from Norwegian ground-measuring stations was used. In addition to solar radiation measurements and estimates, latitude of locations, altitude, solar zenith angle, and clear-sky index was used as inputs to the regression model. To evaluate the robustness of the proposed model, locations from Sweden were used to check the accuracy of the proposed model. The data from Swedish locations were not used in the training of the model. The RFR was trained on a workstation with 16 cores and 64 GB of RAM.

Table 6.4: RMSD, MABD, and MBD of the input data sets and the presented model are shown. The metrics are shown for different geographical locations, including below 65ºN, above 65ºN, coastal, and inland regions. Numbers without parentheses are monthly averaged errors while those in parentheses are daily averaged errors.

	RMSD (Wm^{-2})			MABD (Wm^{-2})			MBD (Wm^{-2})		
	CLARA	**ERA5**	**Model**	**CLARA**	**ERA5**	**Model**	**CLARA**	**ERA5**	**Model**
All sites	9.6 (19.1)	10.2 (26.7)	6.6 (15.7)	6.3 (13.1)	7.0 (16.7)	4.3 (10.2)	-1.6 (-2.0)	3.9 (3.9)	-0.2 (-0.2)
Above 65ºN	9.6 (16.0)	10.1 (26.3)	6.5 (13.7)	6.3 (9.7)	6.9 (14.5)	4.2 (8.2)	-1.6 (-2.9)	3.8 (5.6)	-0.2 (-0.1)
Below 65ºN	9.7 (19.5)	12.7 (26.8)	8.0 (15.9)	6.5 (13.6)	9.4 (17.3)	5.4 (10.5)	-1.8 (-1.8)	5.7 (3.9)	0.1 (-0.1)
Coastal	9.7 (16.7)	10.1 (26.7)	6.6 (14.8)	6.4 (11.4)	7.0 (16.3)	4.3 (9.4)	-1.7 (-1.1)	3.8 (4.9)	-0.2 (0.4)
Inland	8.2 (20.8)	11.2 (26.7)	6.6 (16.4)	5.7 (14.4)	7.9 (17.5)	4.6 (10.8)	-0.6 (-2.6)	4.5 (3.4)	0.1 (-0.4)

Table 6.4 shows the errors in CLARA, ERA5 and the proposed model for Norwegian locations. The model improves the MABD by more than 20%. On monthly averages for all sites, CLARA had an MABD of 6.3 Wm^{-2}, ERA5 had an MABD of 7.0 Wm^{-2}, and the proposed regression model had an MABD of 4.3 Wm^{-2}, which shows a relative improvement of 32% and 39% with respect to CLARA and ERA5. The RMSD of the proposed model was also smaller than CLARA and ERA5, with improvements of 31% and 35%, respectively. However, the bias or MBD was negative for the proposed model as in the case of CLARA. The reason for the negative bias is that CLARA is a more accurate dataset than ERA5; hence, in the regression, more weightage is given to CLARA than ERA5. However, the magnitude of bias in the proposed model is smaller than CLARA. From the bias-variance decomposition of mean squared error ($MSE=RMSD^2$), the variance can be computed as: $Var=RMSD^2-Bias^2$. We can use this to use that the variances of CLARA and ERA5 are very similar, and the variance of the RFR model is less half of these. This proves that the RFR model also provides a large improvement in precision.

Moreover, the R^2 values and the standard deviation (STD) of the Norwegian locations were analyzed as well. Values of the coefficient of determination, R^2, are computed from the ground-measured and model data. The standard deviation is a measure of the spread of the prediction errors around their mean value. Table 6.5 shows the R^2 values and standard deviation for all Norwegian locations, in addition to below 65°N, above 65°N, coastal and inland regions. The standard deviation in Table 6.5 has units of Wm^{-2}, whereas R^2 has no units. For standard deviation, the smaller the value, the better the model estimates and for R^2, the larger the value, the better are the estimates.

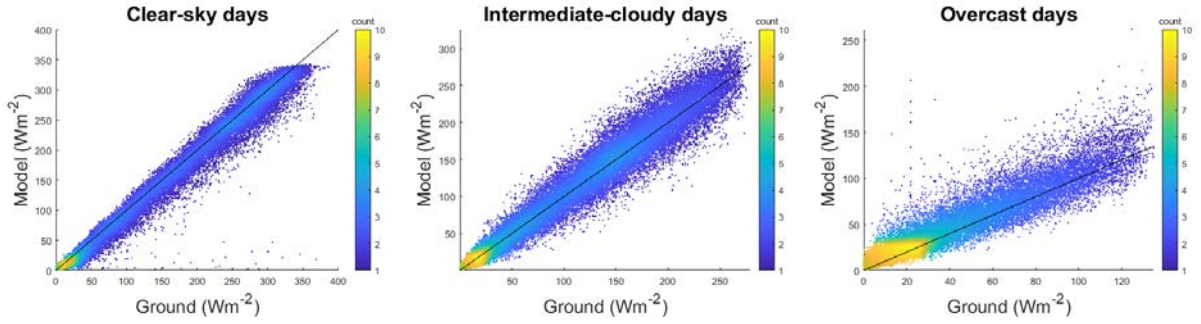
Table 6.5: The R^2 and error standard deviation analysis of CLARA, ERA5, and the proposed RFR model for Norwegian locations is shown here. The RFR model improves the estimates in all types of geographical categories. The unit of the standard deviation (STD) is Wm-2 and R2 is unit-less. Best results are indicated in bold.

	NIBIO sites		**Above 65ºN**		**Below 65ºN**		**Coastal**		**Inland**	
	R^2	**STD**	**R^2**	**STD**	**R^2**	**STD**	**R^2**	**STD**	**R^2**	**STD**
CLARA	0.96	23.8	0.96	18.4	0.95	25.0	0.97	21.1	0.95	25.9
ERA	0.92	26.9	0.89	28.5	0.92	26.7	0.91	27.1	0.92	26.7
RFR (proposed)	**0.97**	**16.0**	**0.97**	**15.3**	**0.97**	**16.1**	**0.97**	**15.3**	**0.97**	**16.5**

The proposed regression model improves the solar radiation estimates at all Norwegian locations. The largest improvements were observed in location above 65°N, although the

differences are small. The proposed model had lower standard deviation than CLARA and ERA5 in all geographical groups.

The error analysis was also performed for locations above 65ºN, below 65ºN, coastal and inland regions. Although the model improved the estimated solar radiation, most of the improvements were seen in coastal regions and regions lying above 65ºN. In addition, a seasonal analysis was also performed on the accuracy of the proposed dataset (see Paper IV). Major improvements were observed in the period of February to July, which evidently are the months that receive largest portion of solar radiation in a year at high latitude locations. One of the shortcomings of the CLARA dataset is the high errors when the solar elevations angles are very low, as in the case of early winter period and late summer period. On the contrary, in these periods ERA5 provides better estimates than CLARA does. The proposed model takes advantage of ERA5 capabilities of improved surface solar radiation estimates at low solar elevation angles and improves the estimates by weighing ERA5 more at these times.



Model	RMSD (Wm^{-2})	MABD (Wm^{-2})	MBD (Wm^{-2})
Clear-sky	17.4	11.3	-6.6
Intermediate cloudy	16.8	11.8	1.7
Overcast	12.8	8.2	5.3

Figure 6.10: Proposed regression model errors under clear-sky, intermediate cloudy and overcast skies. The scatter plots for different sky conditions are also shown. The colored legend bar shows the density of points.

In this study, the sky stratification capability of the proposed data set was studied to assess its performance in different sky conditions. Figure 6.10 show the scatter plots of proposed model in different sky conditions. The method used in sky stratifications used here is the same as shown in the previous section and Paper III.

The proposed model improved the surface solar radiation accuracies in all three sky-categories. Large improvements were observed in clear-sky and intermediate-cloudy categories, while a somewhat small improvement was observed in overcast category.

In the final analysis, solar radiation estimates from the model were evaluated against five Swedish ground-measuring stations. For information on these station refer to Appendix Table C (stations marked with *). As previously explained, the data from Swedish locations was not used in training the regression model. The analysis is shown in Table 6.6, and it can be seen that the model improves the solar radiation estimates in Swedish locations. This robustness test shows that this model can be used to improve solar radiation estimates at high latitude locations.

Table 6.6: The RMSD, MABD, and MBD of the input data sets and the regression model for Swedish locations are listed. These locations were not used in the training of the regression model. Numbers without parentheses are monthly averaged errors while those in parentheses are daily averaged errors.

	RMSD (Wm^{-2})			**MABD (Wm^{-2})**			**MBD (Wm^{-2})**		
	CLARA	**ERA5**	**Model**	**CLARA**	**ERA5**	**Model**	**CLARA**	**ERA5**	**Model**
Kiruna	17.2 (26.6)	7.6 (24.0)	11.0 (18.7)	10.1 (16.6)	4.9 (14.4)	6.8 (11.7)	-7.0 (-8.2)	-2.3 (-2.5)	-5.9 (-6.0)
Luleå	10.6 (24.4)	10.4 (25.1)	5.6 (17.5)	6.9 (14.9)	6.6 (15.3)	3.8 (11.0)	-4.4 (-4.2)	5.1 (4.9)	-2.1 (-2.1)
Umeå	8.3 (16.4)	7.1 (23.0)	5.5 (13.5)	6.1 (11.5)	4.4 (14.2)	3.8 (9.1)	-3.2 (-3.5)	2.0 (2.1)	-2.6 (-2.5)
Stockholm	6.8 (16.4)	7.0 (23.6)	5.9 (14.6)	5.1 (11.5)	4.8 (15.7)	4.5 (10.0)	2.6 (2.5)	3.1 (3.1)	3.9 (4.0)
Göteborg	4.7 (14.9)	9.5 (26.1)	4.8 (14.4)	3.5 (10.5)	7.3 (17.0)	3.7 (9.9)	1.6 (1.8)	6.9 (6.8)	3.0 (2.9)
SMHI locations	10.4 (20.3)	8.4 (24.4)	6.9 (15.9)	6.3 (13.0)	5.6 (15.3)	4.5 (10.3)	-2.1 (-2.3)	2.9 (2.9)	-0.8 (-0.7)

7. Discussion and future work

This chapter presents a discussion on the topics covered in this thesis, appended paper, and provides an overview of the future work.

7.1 Discussion

This research provides an in-depth evaluation of surface solar radiation estimation datasets for high-latitude locations. The solar energy penetration in Norway has been very low when compared to the neighboring countries. One of the hindrances in having higher penetration is the available data and maps for a feasible decision making process. Ground measuring stations are sparse and there are a handful of these stations recording surface solar radiation at high latitude locations. The quality control of the ground-measured data is another important issue as it was observed in this work.

Remotely sensed solar radiation data by satellites provides accurate estimation in mid latitude and equatorial regions, however, at high latitude regions these dataset deteriorate because of the complex viewing angles between terrain, satellites and the Sun. This thesis provides an overview of these available resources for high latitudes along with their accuracies.

Recent studies have shown that the solar radiation estimation from reanalyses has been improving. These dataset provide a valuable support to the satellite datasets, which are currently more accurate than reanalyses. As the ground measuring stations in Norway are located at large distances from each other, reanalyses provide the most reliable and feasible solar radiation estimates for filling gaps in ground-measured and satellite data.

In this thesis, advanced regression method was used to improve the surface solar radiation estimates from satellite and reanalyses datasets. With increasing computing power and sophisticated machine learning algorithms, large datasets are now easier to model. These methods show that with low computing power, large improvements can be made in the available data. In addition to solar radiation, these methods can be used on other renewable energy sources.

7.2 Future work

This thesis provides models and research regarding solar radiation estimation in high-latitude locations. There could be a number of extensions to this works. Some planned research targets are as follows:

- The main solar radiation measurement provider in Norway is NIBIO. This research has found quality control issues in the *in-situ* measurements in the NIBIO data. This data can be quality controlled by using advanced and sophisticated methods. Flags can be introduced for erroneous data, in addition to replacing the erroneous data with the model datasets analyzed in this thesis.
- The regression model presented in Paper IV is more accurate than other available dataset. However, the highest temporal resolution of the proposed data set is limited to daily averages. A future extension of this work includes increasing the temporal and spatial resolution of this data set by using statistical methods.
- The evaluation of the regression model was limited to Scandinavia in this thesis. The data used to train the model was also limited to Norwegian locations. An interesting research extension could be to include data from northern American and Russian regions, so as to have a larger training and testing datasets.
- The datasets analyzed here and the proposed model will be used in performing multiple rooftop solar potential studies by using ArcGIS.
- The new regression based dataset should be used to compute and present a complete solar radiation resource map over the entire Scandinavia and other high latitude regions.
- As shown in this thesis, satellite estimation of solar radiation deteriorates on snow-covered surfaces. A possible research extension is to investigate snow-covered areas through auxiliary data, e.g. IR data from satellites, snow depth data from ERA5 and improve the surface solar radiation on snow-covered surfaces.

8. Summary of conclusions

This chapter summarizes the main conclusions of the research presented in this thesis. These concluding remarks are related to the aims of the thesis presented in Section 1.1 and the knowledge gaps indicated in Section 5.

Modelling surface solar radiation by using meteorological variables (Paper I)

- It was shown that meteorological variables could be used to estimate surface solar radiations in high latitude locations. Moreover, when compared to other such models, the inclusion of relative humidity improves the results. These kind of models can be used at meteorological stations that do not record surface solar radiations.

Comparative analysis of CLARA-A1 and CLARA-A2 (Paper II)

- The CLARA datasets provide surface solar radiation estimates in the Polar Regions. In 2017, the latest version of this dataset called CLARA-A2 was published. A study was performed to assess the improvement of the new edition. It was found that the new edition is more accurate than the CLARA-A1 along with having reduction in the number of missing values. However, the new data points in CLARA-A2 mostly lie on the snow-covered surfaces that have large errors.
- As the northern Scandinavian regions have frequent snow covers in winter, the CLARA-A2 dataset should be used after a proper analysis of land surface properties and biases in solar radiation estimation.

Assessment of satellite and reanalyses datasets for high latitude regions (Paper III)

- A number of satellite and reanalyses provide surface solar radiation estimates at high latitude regions. This analysis showed that surface solar radiation datasets based on satellites provide better estimates than reanalyses.
- Above 65ºN, CLARA delivers the best estimates and below 65ºN, SARAH gives the best estimates. However, the solar radiation estimates from these datasets deteriorate on snow-covered surfaces.

- The newly published reanalysis by ECMWF called the ERA5 provides surface solar radiation estimates on a high temporal resolution. Even though this dataset is not as accurate as satellite dataset, the solar radiation estimates from ERA5 can be used to fill the missing gaps in the monthly mean values of the CLARA datasets.
- Arctic System Reanalysis, which is a polar optimized downscaling of ERA-Interim, was found to have very large errors.

A random forest regression based solar radiation dataset (Paper IV)

- The knowledge gained from preceding studies was used in proposing a dataset based on a random forest regression by using CLARA-A2, ERA5 and auxiliary data. It was found that the proposed model has a considerably improved accuracy compared to CLARA-A2 and ERA5.
- The proposed regression model was trained on 20% data from Norwegian locations. On the Norwegian testing data, substantial improvements were observed. In addition, the same regression model that was trained on Norwegian data was also tested on five Swedish locations with very similar improvements.

Acknowledgements

I would first like to express my deep gratitude to my supervisor, Tobias Boström, for his guidance, enthusiastic encouragement, and useful critiques of this research work. This has been a long and arduous journey, and I found Tobias in all the high and lows with all the supports that I could require. Sincere thanks to Rune Grand Graversen, the co-supervisor of this work. I gained immense knowledge under his guidance and I learned to analyze and present scientific facts with depth and precision. I could not have imagined having better mentors for my PhD study. Special mention goes to Stian Norman Anfinsen, for guiding and working with me in the Machine Learning area. I would like to thank Anthony Paul Doulgeris for proofreading my work and for providing useful writing tips.

I extend my gratitude to Dr Annette Hammer and the Energy Meteorology group at the University of Oldenburg, where I had the exchange during my PhD. I deeply appreciate the long and frequent discussions that I had with Dr Annette until the last day of my stay. I am grateful for the assistance provided by the CM-SAF team, particularly Jörg Trentmann and Karl-Göran Karlsson.

I thank my colleagues at UiT and the Energy and Climate group. I would like to thank my office mate, Kine Solbakken, for being there from the start of this PhD until the end, for being on my side and providing all the support. I would like to thank my friends in Tromsø for making the darkness and a bit more of light, bearable. I would especially like to mention Luigi Tommaso Luppino and Dorota Jozwicki, who filled all these years with smiles and laughter. Your craziness brought a Nirvana to my state of mind.

I must express my very profound gratitude to Daia Irina Georgiana, for providing me with unfailing support and continuous encouragement throughout the process of research and writing this thesis. This accomplishment would not have been possible without you. Thank you.

In the end, I would like to thank the PhD defence committee for reading this thesis.

Bilal Babar
Tromsø, January 2019.

References

1. Bithas K, Kalimeris P. A Brief History of Energy Use in Human Societies. Revisiting the Energy-Development Link: Springer; 2016. p. 5-10.
2. Shepherd DG. Historical development of the windmill: National Aeronautics and Space Administration, Office of Management, Scientific and Technical Information Division; 1990.
3. Fouquet R, Pearson PJ. Seven centuries of energy services: The price and use of light in the United Kingdom (1300-2000). The Energy Journal. 2006:139-77.
4. NOAA. National Centers for Environmental Information date accessed 12-09-2018.
5. Arrhenius S. XXXI. On the influence of carbonic acid in the air upon the temperature of the ground. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. 1896;41(251):237-76.
6. Peterson TC, Connolley WM, Fleck J. The myth of the 1970s global cooling scientific consensus. Bulletin of the American Meteorological Society. 2008;89(9):1325-38.
7. Organization WM. Report of the international conference on the assessment of the role of carbon dioxide and of other greenhouse gases in climate variations and associated impacts. WMO Geneva; 1986.
8. Lorius C, Jouzel J, Ritz C, Merlivat L, Barkov N, Korotkevich YS, et al. A 150,000-year climatic record from Antarctic ice. Nature. 1985;316(6029):591.
9. Bohemen Av, Wittenstein M, Tommila M. Energy Policies of IEA Countries: Norway 2017 Review. 2017.
10. Group B. BP statistical review of world energy. 2018. 2018.
11. BP. BP Statistical Review of World Energy. 2018.
12. REN21. Renewables 2018 Global Status Report, (Paris: REN21 Secretariat). 2018.
13. IRENA. Global Energy Transformation: A roadmap to 2050. International Renewable Energy Agency, Abu Dhabi. 2018;This report is available for download from www.irena.org/publications. For further information or to provide feedback,please contact IRENA at info@irena.org.
14. Holm Ø. Norwegian IEA PVPS Task 1. Press release. 2017.
15. Goswami DY. Principles of solar engineering. **Book**. 2015;3rd Edition.
16. Solanki CS. Solar photovoltaics: Fundamentals, technologies and applications. Book. 2015;3rd Edition.
17. Gueymard CA. Revised composite extraterrestrial spectrum based on recent solar irradiance observations. Solar Energy. 2018;169:434-40.
18. Herron JA, Kim J, Upadhye AA, Huber GW, Maravelias CT. A general framework for the assessment of solar fuel technologies. Energy & Environmental Science. 2015;8(1):126-57.
19. Honsberg C, Bowden S. PV education. ORG(access April-June 2013) http://pveducation org/pvcdrom/properties-of-sunlight/sun-position-calculator. 2014.

20. Aliyu M, Hassan G, Said SA, Siddiqui MU, Alawami AT, Elamin IMJR, et al. A review of solar-powered water pumping systems. 2018;87:61-76.
21. Iqbal M. An introduction to solar radiation: Elsevier; 2012.
22. Lave M, Kleissl J. Optimum fixed orientations and benefits of tracking for capturing solar radiation in the continental United States. Renewable Energy. 2011;36(3):1145-52.
23. Rizk J, Chaiko Y. Solar tracking system: more efficient use of solar panels. World Academy of Science, Engineering and Technology. 2008;41:313-5.
24. Kelly NA, Gibson TL. Improved photovoltaic energy output for cloudy conditions with a solar tracking system. Solar Energy. 2009;83(11):2092-102.
25. Eldin SS, Abd-Elhady M, Kandil H. Feasibility of solar tracking systems for PV panels in hot and cold regions. Renewable Energy. 2016;85:228-33.
26. Kleven Ø, Persson H, Good C, Sulkowski W, Boström T, editors. Solar cells above the Arctic Circle–a comparison between a two-axis tracking system and simulations. Proceedings 24th European Photovoltaic Solar Energy Conference, Hamburg, Germany; 2009.
27. Kipp and Zonen. 2018.
28. Urraca R, Huld T, Gracia-Amillo A, Martinez-de-Pison FJ, Kaspar F, Sanz-Garcia A. Evaluation of global horizontal irradiance estimates from ERA5 and COSMO-REA6 reanalyses using ground and satellite-based data. Solar Energy. 2018;164:339-54.
29. Babar B, Rune G, Tobias B. Solar radiation estimation at high latitudes: Assessment of the CMSAF databases, ASR and ERA5. under-review in Solar Energy. 2018.
30. Urraca R, Gracia-Amillo AM, Koubli E, Huld T, Trentmann J, Riihelä A, et al. Extensive validation of CM SAF surface radiation products over Europe. Remote sensing of environment. 2017;199:171-86.
31. Babar B, Graversen R, Boström T. Evaluating CM-SAF solar radiation CLARA-A1 and CLARA-A2 datasets in Scandinavia. Solar Energy. 2018;170:76-85.
32. Bojanowski JS. Quantifying solar radiation at the earth surface with meteorological and satellite data. 2014.
33. Huld T, Müller R, Gambardella A. A new solar radiation database for estimating PV performance in Europe and Africa. Solar Energy. 2012;86(6):1803-15.
34. Huld T, Pascua IP, Amilla AG. PVGIS 5: New algorithms and features. JRC technical reports,. 2017.
35. Dumortier D, Fontoynont M, Heinemann D, Hammer A, Olseth J, Skartveit A, et al. Satel-Light, a www server which provides high quality daylight and solar radiation data for Western and Central Europe. CIE 24th Session, Warsaw. 1999:277-81.
36. Espinar B, Blanc P, Wald L, Gschwind B, Ménard L, Wey E, et al., editors. HelioClim-3: a near-real time and long-term surface solar irradiance database. Workshop on" Remote Sensing Measurements for Renewable Energy"; 2012.

37. Šúri M, Cebecauer T, editors. SolarGIS: New web-based service offering solar radiation data and PV simulation tools for Europe, North Africa and Middle East. Proceedings of the International Conference on Solar Heating, Cooling and Buildings EUROSUN; 2010.
38. Remund J, Müller SC, editors. Solar radiation and uncertainty information of Meteonorm 7. Proceedings of 26th European Photovoltaic Solar Energy Conference and Exhibition; 2011.
39. Hersbach H, Dee D. ERA5 reanalysis is in production. ECMWF newsletter. 2016;147(7).
40. Meyer R, Hoyer C, Schillings C, Trieb F, Diedrich E, Schroedter M, editors. SOLEMI: A new satellite-based service for high-resolution and precision solar radiation data for Europe, Africa and Asia. Proceedings of the 2003 ISES Solar World Congress, GÃuteborg, Sweden; 2003.
41. Satel-light. Frequency of sunny skies from Satel-light. 2018.
42. Häggmark L, Ivarsson K-I, Gollvik S, Olofsson P-OJTADM, Oceanography. Mesan, an operational mesoscale analysis system. 2000;52(1):2-20.
43. SMHI. STRÅNG - a mesoscale model for solar radiation 2018 [Available from: http://strang.smhi.se/.
44. Øyvind B, Anne L, Løvholm, Sonia L. Resource mapping of solar energy an overview of available data in Norway. Kjeller Vindteknikk

2013;Report KVT/OB/2013/R046.
45. Bengtsson L, Shukla J. Integration of space and in situ observations to study global climate change. Bulletin of the American Meteorological Society. 1988;69(10):1130-43.
46. Trenberth KE, Olson JG. An evaluation and intercomparison of global analyses from the National Meteorological Center and the European Centre for Medium Range Weather Forecasts. Bulletin of the American Meteorological Society. 1988;69(9):1047-57.
47. Xia X, Wang P, Chen H, Liang F. Analysis of downwelling surface solar radiation in China from National Centers for Environmental Prediction reanalysis, satellite estimates, and surface observations. Journal of Geophysical Research: Atmospheres. 2006;111(D9).
48. Linares-Rodríguez A, Ruiz-Arias JA, Pozo-Vázquez D, Tovar-Pescador J. Generation of synthetic daily global solar radiation data based on ERA-Interim reanalysis and artificial neural networks. Energy. 2011;36(8):5356-65.
49. You Q, Sanchez-Lorenzo A, Wild M, Folini D, Fraedrich K, Ren G, et al. Decadal variation of surface solar radiation in the Tibetan Plateau from observations, reanalysis and model simulations. Climate dynamics. 2013;40(7-8):2073-86.
50. Schroeder TA, Hember R, Coops NC, Liang S. Validation of solar radiation surfaces from MODIS and reanalysis data over topographically complex terrain. Journal of Applied Meteorology and Climatology. 2009;48(12):2441-58.

51. Polo J, Wilbert S, Ruiz-Arias JA, Meyer R, Gueymard C, Suri M, et al. Preliminary survey on site-adaptation techniques for satellite-derived and reanalysis solar radiation datasets. Solar Energy. 2016;132:25-37.
52. Jia B, Xie Z, Dai A, Shi C, Chen F. Evaluation of satellite and reanalysis products of downward surface solar radiation over East Asia: Spatial and seasonal variations. Journal of Geophysical Research: Atmospheres. 2013;118(9):3431-46.
53. Trolliet M, Walawender JP, Bourlès B, Boilley A, Trentmann J, Blanc P, et al. Downwelling surface solar irradiance in the tropical Atlantic Ocean: a comparison of re-analyses and satellite-derived data sets to PIRATA measurements. Ocean Science. 2018;14(5):1021-56.
54. Saha S, Moorthi S, Wu X, Wang J, Nadiga S, Tripp P, et al. The NCEP climate forecast system version 2. Journal of Climate. 2014;27(6):2185-208.
55. Dee DP, Uppala SM, Simmons A, Berrisford P, Poli P, Kobayashi S, et al. The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. Quarterly Journal of the royal meteorological society. 2011;137(656):553-97.
56. Kobayashi S, Ota Y, Harada Y, Ebita A, Moriya M, Onoda H, et al. The JRA-55 reanalysis: General specifications and basic characteristics. Journal of the Meteorological Society of Japan Ser II. 2015;93(1):5-48.
57. Onogi K, Tsutsui J, Koide H, Sakamoto M, Kobayashi S, Hatsushika H, et al. The JRA-25 reanalysis. Journal of the Meteorological Society of Japan Ser II. 2007;85(3):369-432.
58. Kanamitsu M, Ebisuzaki W, Woollen J, Yang S-K, Hnilo J, Fiorino M, et al. Ncep–doe amip-ii reanalysis (r-2). Bulletin of the American Meteorological Society. 2002;83(11):1631-44.
59. Rienecker MM, Suarez MJ, Gelaro R, Todling R, Bacmeister J, Liu E, et al. MERRA: NASA's modern-era retrospective analysis for research and applications. Journal of climate. 2011;24(14):3624-48.
60. Compo GP, Whitaker JS, Sardeshmukh PD, Matsui N, Allan RJ, Yin X, et al. The twentieth century reanalysis project. Quarterly Journal of the Royal Meteorological Society. 2011;137(654):1-28.
61. Michalakes J. Design of a next-generation regional weather research and forecast model. Argonne National Lab., IL (US); 1999.
62. Fritz S, Rao PK, Weinstein MJJotAS. Satellite measurements of reflected solar energy and the energy received at the ground. 1964;21(2):141-51.
63. Cano D, Monget J-M, Albuisson M, Guillard H, Regas N, Wald L. A method for the determination of the global solar radiation from meteorological satellites data. Solar energy. 1986;37(1):31-9.
64. Ineichen P, Barroso CS, Geiger B, Hollmann R, Marsouin A, Mueller RJIJoRS. Satellite Application Facilities irradiance products: hourly time step comparison and validation over Europe. 2009;30(21):5549-71.

65. Carlund T. Upgrade of SMHI's meteorological radiation network 2006-2007: Effects on direct and global solar radiation: SMHI; 2011.
66. NIBIO. Les mer om LMT 2018 [Available from: http://lmt.bioforsk.no/about.
67. Long CN, Dutton EG. BSRN Global Network recommended QC tests, V2. x. 2010.
68. Urraca R, Gracia-Amillo AM, Huld T, Martinez-de-Pison FJ, Trentmann J, Lindfors AV, et al. Quality control of global solar radiation data with satellite-based products. Solar Energy. 2017;158:49-62.
69. Ångström AJQJotRMS. Solar and terrestrial radiation. Report to the international commission for solar research on actinometric investigations of solar and atmospheric radiation. 1924;50(210):121-6.
70. Supit I, Van Kappel RJSE. A simple method to estimate global radiation. 1998;63(3):147-60.
71. Abraha M, Savage MJA, Meteorology F. Comparison of estimates of daily solar radiation from air temperature range for application in crop simulations. 2008;148(3):401-16.
72. Weiss A, Hays CJJA, meteorology F. Simulation of daily solar irradiance. 2004;123(3-4):187-99.
73. Wu G, Liu Y, Wang TJEc, management. Methods and strategy for modeling daily global solar radiation with measured meteorological data–A case study in Nanchang station, China. 2007;48(9):2447-52.
74. Hargreaves GL, Hargreaves GH, Riley JPJJoI, Engineering D. Irrigation water requirements for Senegal River basin. 1985;111(3):265-75.
75. Bristow KL, Campbell GSJA, meteorology f. On the relationship between incoming solar radiation and daily maximum and minimum temperature. 1984;31(2):159-66.
76. Hargreaves GH, Samani ZAJJotI, Division D. Estimating potential evapotranspiration. 1982;108(3):225-30.
77. Karlsson K-G, Riihelä A, Müller R, Meirink J, Sedlar J, Stengel M, et al. CLARA-A1: the CM SAF cloud, albedo and radiation dataset from 28 yr of global AVHRR data. Atmospheric Chemistry & Physics Discussions. 2013;13(1).
78. Trentmann J, Kothe S. Surface Radiation Products

Product User Manual. CM SAF Cloud, Albedo, Radiation dataset,

. 2016;SAF/CM/DWD/PUM/GAC/RAD(2.1).
79. Haase J, Calais E, Talaya J, Rius A, Vespe F, Santangelo R, et al. The contributions of the MAGIC project to the COST 716 objectives of assessing the operational potential of ground-based GPS meteorology on an international scale. 2001;26(6-8):433-7.
80. Karlsson K-G, Anttila K, Trentmann J, Stengel M, Meirink JF, Devasthale A, et al. CLARA-A2: the second edition of the CM SAF cloud and radiation data record from 34 years of global AVHRR data. 2017;17(9):5809-28.

81. Karlsson K-G, Anttila K, Trentmann J, Stengel M, Meirink JF, Devasthale A, et al. CLARA-A2: the second edition of the CM SAF cloud and radiation data record from 34 years of global AVHRR data. Atmospheric Chemistry and Physics. 2017;17(9):5809.

82. Pfeifroth U, Kothe S, Müller R, Trentmann J, Hollmann R, Fuchs P, et al. Surface Radiation Data Set–Heliosat (SARAH)–Edition 2, Satellite Application Facility on Climate Monitoring. 2017.

83. Kothe S, Pfeifroth U, Cremer R, Trentmann J, Hollmann R. A satellite-based sunshine duration climate data record for Europe and Africa. Remote Sensing. 2017;9(5):429.

84. Riihelä A, Carlund T, Trentmann J, Müller R, Lindfors AV. Validation of CM SAF surface solar radiation datasets over Finland and Sweden. Remote Sensing. 2015;7(6):6663-82.

85. Hammer A, Heinemann D, Hoyer C, Kuhlemann R, Lorenz E, Müller R, et al. Solar energy assessment using remote sensing technologies. Remote Sensing of Environment. 2003;86(3):423-32.

86. Mueller R, Behrendt T, Hammer A, Kemper A. A new algorithm for the satellite-based retrieval of solar surface irradiance in spectral bands. Remote Sensing. 2012;4(3):622-47.

87. Pfeifroth U, Sanchez-Lorenzo A, Manara V, Trentmann J, Hollmann R. Trends and Variability of Surface Solar Radiation in Europe Based On Surface-and Satellite-Based Data Records. Journal of Geophysical Research: Atmospheres. 2018;123(3):1735-54.

88. Müller R, Pfeifroth U, Träger-Chatterjee C, Cremer R, Trentmann J, Hollmann R. Surface solar radiation data set-Heliosat (SARAH). Climate Monitoring Satellite Application Facility: Darmstadt, Germany; 2015.

89. ECMWF. ERA5 data documentation. 2018.

90. Zhao L, Lee X, Liu S. Correcting surface solar radiation of two data assimilation systems against FLUXNET observations in North America. Journal of Geophysical Research: Atmospheres. 2013;118(17):9552-64.

91. Bromwich DH, Hines KM, Bai L-S, editors. Arctic system reanalysis. Third World Climate Research Programme International Conference on Reanalysis Tokyo, Japan, January; 2008.

92. Bromwich D, Kuo YH, Serreze M, Walsh J, Bai LS, Barlage M, et al. Arctic system reanalysis: call for community involvement. 2010;91(2):13-4.

93. Bromwich D, Wilson A, Bai L, Liu Z, Barlage M, Shih C-F, et al. The Arctic System Reanalysis, Version 2. 2018;99(4):805-28.

94. Bromwich DH, Wilson AB, Bai LS, Moore GW, Bauer PJQJotRMS. A comparison of the regional Arctic System Reanalysis and the global ERA-Interim Reanalysis for the Arctic. 2016;142(695):644-58.

95. Skamarock WC, Klemp JB, Dudhia J, Gill DO, Barker DM, Wang W, et al. A description of the advanced research WRF version 2. National Center For Atmospheric Research Boulder Co Mesoscale and Microscale …; 2005.

96. Moradi IJE. Quality control of global solar radiation using sunshine duration hours. 2009;34(1):1-6.

97. Younes S, Claywell R, Muneer TJE. Quality control of solar radiation data: Present status and proposed new approaches. 2005;30(9):1533-49.
98. Journée M, Bertrand CJSE. Quality control of solar radiation data within the RMIB solar measurements network. 2011;85(1):72-86.
99. Davies J, McKay DJSE. Evaluation of selected models for estimating solar radiation on horizontal surfaces. 1989;43(3):153-68.
100. König-Langlo G, Sieger R, Schmithüsen H, Bücker A, Richter F, Dutton EJWR. Baseline Surface Radiation Network (BSRN)—Update of the technical plan for BSRN data management. 2013.
101. Liaw A, Wiener MJRn. Classification and regression by randomForest. 2002;2(3):18-22.
102. Breiman LJMl. Random forests. 2001;45(1):5-32.
103. Segal MR. Machine learning benchmarks and random forest regression. 2004.
104. Bylander TJML. Estimating generalization error on two-class datasets using out-of-bag estimates. 2002;48(1-3):287-97.
105. Siroky DSJSS. Navigating random forests and related advances in algorithmic modeling. 2009;3:147-63.
106. Luppino LT, Bianchi FM, Moser G, Anfinsen SN. Remote sensing image regression for heterogeneous change detection. arXiv preprint arXiv:180711766. 2018.
107. Schwandt M, Chhatbar K, Meyer R, Fross K, Mitra I, Vashistha R, et al. Development and test of gap filling procedures for solar radiation data of the Indian SRRA measurement network. 2014;57:1100-9.
108. Hollmann R, Mueller R, Gratzki A. CM-SAF surface radiation budget: First results with AVHRR data. Advances in Space Research. 2006;37(12):2166-71.
109. Posselt R, Mueller R, Stöckli R, Trentmann J. Spatial and temporal homogeneity of solar surface irradiance across satellite generations. Remote Sensing. 2011;3(5):1029-46.
110. Müller R, Pfeifroth U, Träger-Chatterjee C, Trentmann J, Cremer R. Digging the METEOSAT treasure—3 decades of solar surface radiation. Remote Sensing. 2015;7(6):8067-101.
111. Bojanowski JS, Vrieling A, Skidmore AK. A comparison of data sources for creating a long-term time series of daily gridded solar radiation for Europe. Solar Energy. 2014;99:152-71.
112. Sanchez-Lorenzo A, Enriquez-Alonso A, Wild M, Trentmann J, Vicente-Serrano SM, Sanchez-Romero A, et al. Trends in downward surface solar radiation from satellites and ground observations over Europe during 1983–2010. Remote Sensing of Environment. 2017;189:108-17.
113. Schulz J, Albert P, Behr H-D, Caprion D, Deneke H, Dewitte S, et al. Operational climate monitoring from space: the EUMETSAT Satellite Application Facility on Climate Monitoring (CM-SAF). Atmospheric Chemistry and Physics. 2009;9(5):1687-709.
114. Posselt R, Mueller R, Stöckli R, Trentmann J. Remote sensing of solar surface radiation for climate monitoring—The CM-SAF retrieval in international comparison. Remote Sensing of Environment. 2012;118:186-98.

115. Cristóbal J, Anderson M. Validation of a Meteosat Second Generation solar radiation dataset over the northeastern Iberian Peninsula. Hydrology and Earth System Sciences. 2013;17(1):163-75.
116. Journée M, Stöckli R, Bertrand C. Sensitivity to spatio-temporal resolution of satellite-derived daily surface solar irradiation. Remote sensing letters. 2012;3(4):315-24.
117. Amillo AG, Huld T, Müller R. A new database of global and direct solar radiation using the eastern meteosat satellite, models and validation. Remote sensing. 2014;6(9):8165-89.
118. Alexandri G, Georgoulias A, Meleti C, Balis D, Kourtidis K, Sanchez-Lorenzo A, et al. A high resolution satellite view of surface solar radiation over the climatically sensitive region of Eastern Mediterranean. Atmospheric Research. 2017;188:107-21.
119. Urraca R, Martinez-de-Pison E, Sanz-Garcia A, Antonanzas J, Antonanzas-Torres F. Estimation methods for global solar radiation: Case study evaluation of five different approaches in central Spain. Renewable and Sustainable Energy Reviews. 2017;77:1098-113.
120. Žák M, Mikšovský J, Pišoft P. CMSAF radiation data: New possibilities for climatological applications in the Czech Republic. Remote Sensing. 2015;7(11):14445-57.
121. Posselt R, Müller R, Trentmann J, Stockli R, Liniger MA. A surface radiation climatology across two Meteosat satellite generations. Remote sensing of environment. 2014;142:103-10.
122. Hakuba MZ, Folini D, Sanchez-Lorenzo A, Wild M. Spatial representativeness of ground-based solar radiation measurements—Extension to the full Meteosat disk. Journal of Geophysical Research: Atmospheres. 2014;119(20):11,760-11,71.
123. Bennett K. Milankovitch cycles and their effects on species in ecological and evolutionary time. Paleobiology. 1990;16(1):11-21.
124. Bird RE, Hulstrom RL. Simplified clear sky model for direct and diffuse insolation on horizontal surfaces. Solar Energy Research Inst., Golden, CO (USA); 1981.
125. Smith CJ, Bright JM, Crook R. Cloud cover effect of clear-sky index distributions and differences between human and automatic cloud observations. Solar Energy. 2017;144:10-21.
126. Widén J, Shepero M, Munkhammar J. On the properties of aggregate clear-sky index distributions and an improved model for spatially correlated instantaneous solar irradiance. Solar Energy. 2017;157:566-80.
127. Mueller R, Träger-Chatterjee C. Brief accuracy assessment of aerosol climatologies for the retrieval of solar surface radiation. Atmosphere. 2014;5(4):959-72.
128. Polo J, Antonanzas-Torres F, Vindel J, Ramirez L. Sensitivity of satellite-based methods for deriving solar radiation to different choice of aerosol input and models. Renewable energy. 2014;68:785-92.
129. Kishore P, Ratnam MV, Namboothiri S, Velicogna I, Basha G, Jiang J, et al. Global (50 S–50 N) distribution of water vapor observed by COSMIC GPS RO: Comparison with GPS radiosonde, NCEP, ERA-Interim, and JRA-25 reanalysis data sets. Journal of Atmospheric and Solar-Terrestrial Physics. 2011;73(13):1849-60.

Appendix

Table A: Information about the site locations from Norway used in the thesis and appended papers. The table shows the coordinates of ground measuring stations along with their altitudes, and land type. Paper III and IV.

	Station	**Latitude**	**Longitude**	**Altitude**	**Land type**
1	Holt	69.65	18.91	12	Coastal
2	Sortland	68.65	15.28	14	Coastal
3	Vågønes	67.28	14.45	26	Coastal
4	Tjøtta	65.83	12.43	10	Coastal
5	Skogmo	64.51	12.02	32	Inland
6	Rissa	63.59	9.97	23	Coastal
7	Kvithamar	63.49	10.88	28	Inland
8	Skjetlein	63.34	10.3	44	Coastal
9	Surnadal	62.98	8.69	5	Inland
10	Tingvoll	62.91	8.19	23	Coastal
11	Fåvgang	61.46	10.19	184	Inland
12	Fureneset	61.29	5.04	12	Coastal
13	Gausdal	61.22	10.26	375	Inland
14	Løken	61.12	9.06	527	Inland
15	Ilseng	60.8	11.2	182	Inland
16	Kise	60.77	10.81	129	Inland
17	Apelsvoll	60.7	10.87	262	Inland
18	Hønefoss	60.14	10.27	126	Inland
19	Årnes	60.13	11.39	162	Inland
20	Etne	59.66	5.95	8	Inland
21	Ås	59.66	10.78	94	Inland
22	Bø	59.42	9.03	105	Inland
23	Rakkestad	59.39	11.39	102	Inland
24	Ramnes	59.38	10.24	39	Coastal
25	Tomb	59.32	10.81	12	Coastal
26	Gjerpen	59.23	9.58	41	Coastal
27	Hjelmeland	59.23	6.15	43	Inland
28	Tjølling	59.05	10.13	19	Coastal
29	Særheim	58.76	5.65	90	Coastal
30	Landvik	58.34	8.52	10	Coastal
31	Lyngdal	58.13	7.05	4	Inland

Table B: List of years not included in Paper III and IV.

	Station	Years having more than 5% missing data	Years failing Long and Dutton test	Years having operational error (snow/frost/ shading/soiling)	Years having equipment error
1	Holt	2001,2002,2006,2007,2008,2010	2013		2000
2	Sortland	2000,2006,2007,2010,2013			
3	Vågønes	2006,2007		2002	
4	Tjøtta	2006,2007			2008, 2012
5	Skogmo	2006,2007,2008,2015		2011	2013, 2014
6	Rissa	2006,2007	2000		
7	Kvithamar	2006,2007,2013			
8	Skjetlein	2006,2007	2000		
9	Surnadal	2006,2007,2014			
10	Tingvoll	2006,2007,2012			
11	Fåvang	2006,2007			2001
12	Fureneset	2006,2007,2011,2012			
13	Gausdal	2006,2007,2009			2015
14	Løken	2006,2007			
15	Ilseng	2006,2007,2004	2000	2009	
16	Kise	2002,2006,2007,2015		2013	
17	Apelsvoll	2006,2007		2002,2003,2004	2009
18	Hønefoss	2006,2007	2000		
19	Årnes	2006,2007			
20	Etne	2006,2007		2004,2012	
21	Ås	2006,2007			
22	Bø	2000,2006,2007			
23	Rakkestad	2006,2007			
24	Ramnes	2006,2007		2009	
25	Tomb	2006,2007	2009		
26	Gjerpen	2006,2007,2015			
27	Hjelmeland	2006,2007			2002, 2015
28	Tjølling	2006,2007,2008,2014		2012,2015	2009, 2010
29	Særheim	2000,2006,2007			
30	Landvik	2006,2007		2005,2010,2014,2015	
31	Lyngdal	2006,2007	2001		

Table C: Information about the site locations from Sweden used in the Paper II and IV. Location marked with (*) were used in Paper IV. The table shows the coordinates of ground measuring stations along with their altitudes, land type, and years not included in the study.

	Sweden	**Latitude**	**Longitude**	**Altitude**	**Land Cover**	**Years not included**
1	Kiruna*	67.83	20.43	408	Sparse forest	N.A
2	Luleå*	65.55	22.13	17	Coastal/archipelago	N.A
3	Umeå*	63.82	20.25	10	rural	N.A
4	Borlange	60.48	15.43	140	Urban/forest	N.A
5	Stockholm*	59.35	18.07	30	Coastal/archipelago	1998
6	Goteborg*	57.70	12.00	5	Coastal	N.A
7	Lund	55.71	13.21	73	Urban	N.A

Paper I [redacted]

Renew. Energy Environ. Sustain. **1**, 34 (2016)

DOI: 10.1051/rees/2016048

Renewable Energy and Environmental Sustainability

Available online at:
www.rees-journal.org

RESEARCH ARTICLE OPEN ACCESS

Estimating solar irradiation in the Arctic

Bilal Babar* and Tobias Boström

Department of Physics and Technology, The Arctic University of Norway, UiT, Mailbox 6050, Langnes, 9037 Tromsø, Norway

Abstract. Solar radiation data plays an important role in pre-feasibility studies of solar electricity and/or thermal system installations. Measured solar radiation data is scarcely available due to the high cost of installing and maintaining high quality solar radiation sensors (pyranometers). Indirect measured radiation data received from geostationary satellites is unreliable at latitudes above 60 degrees due to the resulting flat viewing angle. In this paper, an empirical method to estimate solar radiation based on minimum climatological data is proposed. Eight sites in Norway are investigated, all of which lie above 60 N. The estimations by the model are compared to the ground measured values and a correlation coefficient of 0.88 was found while over all percentage error was −1.1%. The proposed models is 0.2% efficient on diurnal and 10.8% better in annual estimations than previous models.

1 Introduction

Solar radiation data is required when designing active or passive solar installations. Information about solar radiation is also widely used in agriculture, forestry and biological processes [1]. In this study, the emphasis is on the active solar installations in northern Norway and the Arctic. Solar radiation is not an easily obtained quantity, even though it is of great importance. In the case of northern Europe, solar irradiation data is scarcely available. One of the main reasons is the unavailability of weather stations having pyranometers, and that data from geostationary satellites is not very accurate because of the flat viewing angle. In this study, we present an empirical method to calculate the solar irradiation based on only temperature and relative humidity recordings.

The most straightforward method to measure solar radiation would be the installation of pyranometers, but there are two main limiting constraints in this approach. The first being the high cost of the equipment, and secondly, the regular maintenance. Due to these constraints all over the world and especially in northern Norway, such equipment is often not installed even at the weather stations, set up and maintained by the Norwegian Meteorological Institute. Globally, the percentage of weather stations recording the solar radiation is small, roughly 10%, as compared to the stations recording other climatological quantities like temperature, precipitation, humidity, etc. while the ratio of weather station recording short wave solar radiation to stations recording temperature is 1:500 [1]. In the context of Norway, Bioforsk and Meteorologisk Institut have 70 high quality weather stations recording the radiation. In addition, Energinettet has 32 station installed with pyranometers (low quality recordings, non-ISO9060 compliant) and Norwegian Radiation Protection Authority has 10 stations in Norway [2]. Of the 1044 weather stations in Norway [3], only 112 stations provide radiation data. Online resources are available but they also do not cover the area of Scandinavia thoroughly, for example, PVGIS is having only one station of solar radiation from Norway, while Meteonorm is having three Norwegian station in their database of 1200 worldwide stations [2]. Satellites can be used to estimate solar radiation, but above 60 degrees north the estimations from satellites are not reliable because of the flat viewing angle. Consequently there is a need for finding the solar irradiation quantity using methods for plus 60 degree latitudes. We propose to use an empirical model that can calculate solar irradiation based on only temperature and mean relative humidity as input data.

2 Estimation of solar radiation

Analytical, stochastic, empirical and artificial neural network models have been used in the past for the estimation of solar irradiation [4,5]. In reference [6], the author used satellite images to calculate the ground solar radiation through heliosat, a solar radiation estimation method based on geostationary satellites. The modelling of such a system is very difficult and the required information is most of the times incomplete. Stochastic models are used in [7] to estimate the solar radiation, but because of the linear property of such models, they cannot produce good enough results, as the behavior of solar

* e-mail: bilal.babar@uit.no

radiation, especially in the presence of clouds, is non-linear. Artificial neural network are very competitive in estimating solar radiation. Authors in [8–10] have used this AI technique to estimate the solar radiation. The problem with this technique is the higher computational power required in solving the problem, and secondly, the results are not precise when the area between the observation points is large. Such models do not take into account the regional factors involved in the variation of the solar radiation.

Empirical models for the estimation of solar radiation exists since long. In 1924, one of the first models for such estimations was proposed [11]. With this model, there is always a need to calculate two coefficients, which vary for different areas [12]. In reference [13], the authors showed the dependency of temperature and solar radiation on the evapotranspiration of an area. The proposed model is based on the model from [13], but instead of only temperature difference, the model takes into account the effect of relative humidity as well.

3 Methodology

The Arctic poses a unique problem when it comes to estimating solar radiation as the length of sunlight hour's changes very rapidly, from the sun being below the horizon during two winter months to 24-h sunlight during the summer months. For the estimation of solar radiation, equations from [13] can be used. A general form of the equation is given below:

$$R_s = \mathrm{KT} \times R_e (T_{\max} - T_{\min})^{0.5}, \tag{1}$$

where R_s is the estimated radiation, R_e is the extraterrestrial radiation, KT is constant, $T_{\max}$ and $T_{\min}$ are the maximum and minimum temperatures. In such a model, the global horizontal solar radiation is estimated by the recorded levels of maximum and minimum temperatures on a particular day. The value of constant KT varies from 0.162 for interior regions and 0.19 for coastal regions. The main shortcoming of such a model is that it does not take into account the effect of clouds. When observing radiation, clouds may be the biggest affecting factor, and the variation in the radiation caused by clouds is very rapid and could be at a large scale.

4 Proposed model

In this paper, we propose a novel method to estimate the solar radiation. The proposed method is based on the model given in [13], but instead of using only temperature difference, this method uses the relative humidity as well. By using relative humidity as an extra variable, this system becomes more robust and efficient. In addition, the radiation effect on humidity is twice that of temperature [14]. A critical value of relative humidity results in the cloud formation, which increases from zero at some specified relative humidity, to overcast when relative humidity is 100%. It becomes evident that for an overcast day the estimation model from equation (1) could be improved by taking in to account the relative humidity (a commonly recorded meteorological variable). This model performs relatively better on days having clouds. The equation used in this study is shown below:

$$R_s = 0.04 \times R_e \times (T_{\max} - T_{\min}) + \mathrm{CT} \times R_e \times (\mathrm{RH})^{0.27}. \tag{2}$$

In equation (2), R_s is the estimated horizontal global solar radiation, R_e is the extraterrestrial solar radiation, RH is daily averaged ground-measured relative humidity, $T_{\min}$ and $T_{\max}$ are the minimum and maximum temperatures, respectively. CT is a constant, which varies geographically. The parameter R_e limits the estimated values of the global radiation to certain levels. In the Arctic, between the months of November and January when there is no light, the value of extraterrestrial radiation is zero, driving the estimated curve also to zero. The following equation was used to calculate the extraterrestrial radiation [15].

$$R_e = \frac{24}{\pi} R_{\mathrm{sc}} \left(1 + 0.033 \times \cos\frac{360 \times P}{365}\right) \times \cos(\varphi) \times \cos(\delta) \times \sin(\mathrm{hs}) + \left(\frac{2 \times \pi \times \mathrm{hs}}{360}\right) \times \sin(\varphi) \times \sin(\delta), \tag{3}$$

where R_e is the extraterrestrial radiation, R_{sc} is the solar constant with a value of 1.366 kW/m^2, P is the day number from 1 to 365 (366 leap), φ is the latitude of the area, δ is the declination angle and hs is the hour angles of sunrise and sunset.

5 Results and discussion

In the literature, many types of evaluation techniques are used for finding the accuracy and precision of empirical models. When estimating solar radiation, root mean square error was found to be the most widely used parameter. Other parameters such as standard deviation, mean bias error, mean absolute error and mean square error are also used to find the accuracy of models. However, in reference [16], it is suggested that for such empirical models, root mean square error may results in a higher value if there are a few high values in the sample, while mean bias errors can cancel out if negative and positive biases are present.

In this study, the evaluation of the proposed model was checked with four statistical indices: normalized root mean square error (RMSE), *t*-statistic (*t*-stat), yearly percentage error (YPE) and correlation coefficient (Corr). The model is further evaluated by correlating all the observed and calculated values and plotting the data on a scatter plot.

The proposed model was tested on eight sites in Norway for a period of 10 years. The sites were Tromso, Bodo, Sortland, Tingvoll, Pasvik, Overhalla, Gausdal and Etne. The data for these sites was taken from Bioforsk and all the sites are located at latitude higher than 60 degrees

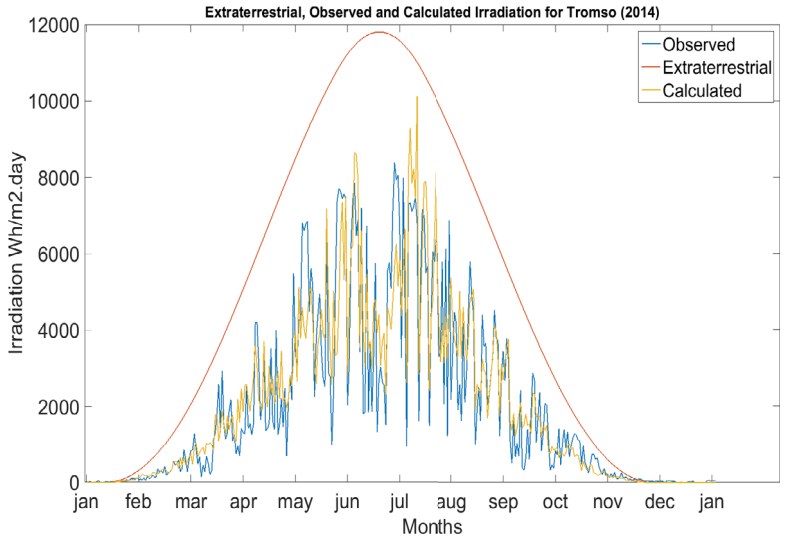
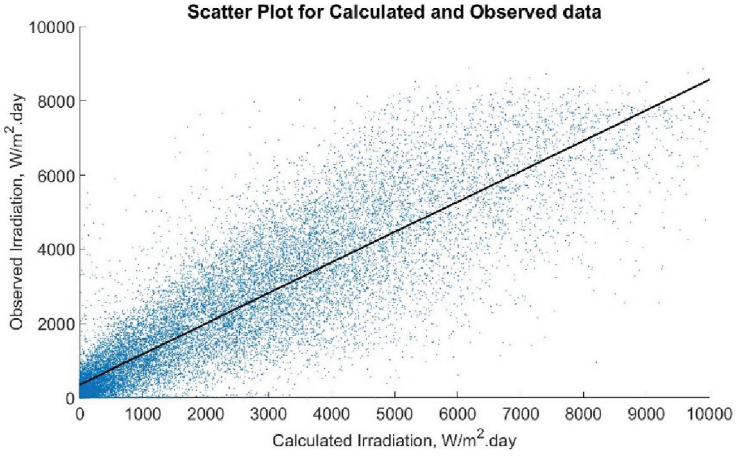
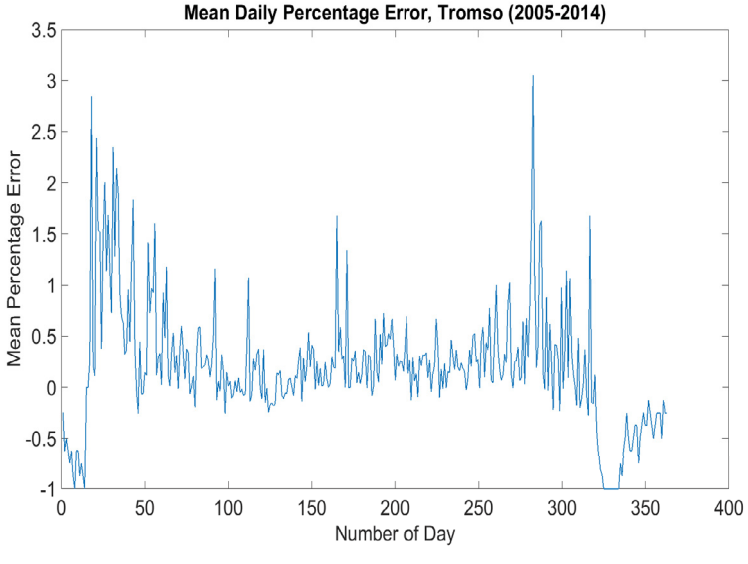


Fig. 1. Observed, calculated and extraterrestrial radiation.



(a)



(b)

Fig. 2. (a) Scatter plot of the data for calculated and observed values, the correlation coefficient for the model is 0.88. (b) Mean percentage error for Tromso, average over 2005–2014.

Table 1. Statistical performance parameters for the sites.

		2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Tromso	RMSE	0.10	0.11	0.11	–	0.10	–	0.09	0.11	0.11	0.09
	t-Stat	0.52	2.5	2.1	–	1.8	–	0.95	0.57	2.2	1.2
	YPE	1.35	8.41	4.8	–	4.87	–	1.84	1.87	6.7	3.6
	Corr	0.9	0.83	0.93	–	0.88	–	0.90	0.88	0.85	0.91
Bodo	RMSE	0.10	0.20	–	0.09	0.11	0.11	0.10	0.09	0.10	–
	t-Stat	0.36	5.5	–	2.1	1.4	0.87	1.22	1.02	3.5	–
	YPE	0.9	22.9	–	4.27	3.31	2.22	3.14	2.3	9.03	–
	Corr	0.89	0.74	–	0.93	0.9	0.88	0.89	0.91	0.90	–
Sortland	RMSE	0.09	0.11	0.15	0.10	0.10	0.14	0.09	0.12	–	0.09
	t-Stat	3.2	3.7	1.6	2.8	0.53	0.52	0.98	1.2	–	2.8
	YPE	7.3	10.9	5.14	6.05	1.07	1.82	2.53	3.16	–	6.9
	Corr	0.91	0.86	0.85	0.92	0.93	0.80	0.90	0.89	–	0.91
Tingvoll	RMSE	–	0.09	0.06	0.09	0.09	–	–	0.10	0.12	0.11
	t-Stat	–	2.9	2.3	1.3	2.7	–	–	1.5	2.3	5.05
	YPE	–	7.28	5.83	2.55	5.5	–	–	3.5	5.51	12.6
	Corr	–	0.90	0.91	0.93	0.92	–	–	0.90	0.90	0.91
Pasvik	RMSE	0.11	–	0.33	0.11	0.10	0.09	0.09	0.10	0.10	0.08
	t-Stat	0.82	–	0.48	0.61	1.8	1.8	0.17	1.25	1.4	1.7
	YPE	2.64	–	3.43	1.6	5.25	6.15	0.3	3.19	3.77	4.4
	Corr	0.86	–	0.16	0.87	0.87	0.85	0.92	0.89	0.89	0.92
Overhala	RMSE	–	0.08	0.05	0.09	0.08	0.11	0.08	0.07	0.09	0.10
	t-Stat	–	7.7	2.1	1.04	2.1	2.5	0.3	0.87	5.2	0.42
	YPE	–	15.1	8.76	2.67	6.2	7.9	0.7	2.13	15.4	1.1
	Corr	–	0.92	0.80	0.90	0.87	0.85	0.92	0.91	0.91	0.90
Gausdal	RMSE	0.09	0.08	0.27	0.10	0.08	0.06	0.11	0.10	–	0.09
	t-Stat	5.59	5.06	2.98	4.7	5.4	3.7	3.5	1.14	–	0.36
	YPE	13.2	12.6	26.1	10.9	13.8	17.5	10.8	2.5	–	0.7
	Corr	0.92	0.93	0.07	0.93	0.91	0.77	0.85	0.90	–	0.93
Etne	RMSE	–	0.09	0.10	0.09	0.10	0.12	0.10	0.09	0.12	0.10
	t-Stat	–	3.5	1.6	2.09	2.7	5.4	2.05	1.8	3.05	1.4
	YPE	–	7.67	4.49	4.38	5.9	11.9	5.06	3.9	7.1	3.3
	Corr	–	0.91	0.89	0.93	0.91	0.90	0.88	0.92	0.88	0.90

north. Sites were selected on the basis that they provide solar radiation recorded by a pyranometer, so that after using the empirical model a correlation could be made for evaluation. Constant CT in equation (2) was found by regressing one year data from the data sets. For areas under consideration, Etne, Overhalla and Pasvik were having CT value of 0.001 while all other areas were having a CT value of 0.04. For the credibility of the model it is very important that the model performs well with the same constants when data from other data bases is used. With 0.04 constant for Tromso, model was checked with data from the weather station of the University of Tromso and similar results were obtained. In addition to the application on higher latitudes, it is expected that the model could be used at almost any place after tuning the constants.

As a comparison, proposed model performs better than the model in equation (1). The average t statistic value for our model is 1.4 as compared to 5.5 from equation (1). The daily average percentage error is improved by 0.2%, while yearly average percentage error is improved by 10.8%. In Figure 1, a yearly plot of radiation is shown. The observed, extraterrestrial and calculated values are daily figures. It can be seen that the estimated values are very close to the observed values of radiation.

In Figure 2(a), a scatter plot is shown for all the eight sites over the 10 years period. A very good positive correlation of 0.88 was found in this case. Furthermore, in Figure 2(b), a graph of daily average errors for Tromso is shown. The percentage errors are calculated for each year and an average was taken for the 10 years period. Negative error values in Figure 2(b) can be observed in the start and end of the year. It is because of the polar night observed at higher latitudes. After and before these negative means there is a high positive mean which is because of the low solar latitude, daily values of irradiation are very low in these days of the year. Both the scatter plot and the daily average error shows a promising result for the model.

In Table 1, we have shown the error statistics for all the sites. These statistics were evaluated for each year separately. The table gives a complete overview of the performance of the model. YPE is in percentage while t-stat, Corr and RMSE (normalized) is unit less. For all the parameters accept correlation coefficient, lower the value better the models performance, while for Corr, the closer to 1 the better is the models performance. The years for which the data was not available were omitted.

6 Conclusion

In this paper, we presented a novel method to calculate the global solar radiation on horizontal surface by using minimum climatological data. For calculating solar radiation, only temperature difference and relative humidity values were used. The models performance was checked on eight sites in Norway. The performance of the model was evaluated through four statistical measures and the results obtained are acceptable, having a correlation coefficient of 0.88 and an overall percentage error of −1.1%. The daily error values of the model are also quite stable where most of the values are lying below 4%.

References

1. V. Badescu, *Modeling solar radiation at the earth's surface: recent advances* (Springer, 2008), ISBN: 978-3-540-77454-9 (Print) 978-3-540-77455-6 (Online)
2. Ø. Byrkjedal, A.N. Løvholm, S. Lileo, Resource mapping of solar energy an overview of available data in Norway report, Kjeller Vindteknikk KVT/OB/2013/R046 (2013)
3. M.A.J. Kvasnes, T. Storaas, H.C. Pedersen, S. Bjørk, E.B. Nilsen, Spatial dynamics of Norwegian tetraonid populations, Ecol. Soc. Jpn. **25**, 367 (2009)
4. G. Mihalakakou, M. Santamouris, D.N. Asimakopoulos, The total solar radiation time series simulation in Athens using neural networks, Theor. Appl. Climatol. **66**, 185 (2000)
5. F.S. Tymvios, C.P. Jacovides, S.C. Michaelides, C. Scouteli, Comparative study of Angstrom's and artificial networks' methodologies in estimating global solar radiation, Sol. Energy **78**, 752 (2005)
6. K. Dagestad, Estimating global radiation at ground level from satellite images, Ph.D. thesis, Department of Meteorology, University of Bergen, Norway, 2005
7. A. Zeroual, M. Ankrim, A.S. Wilkinson, Stochastic modeling of daily global solar radiation measured in Marrakesh Morocco, Renew. Energy **6**, 787 (1995)
8. M. Mohandes, S. Rehman, T.O. Halawani, Estimation of global solar radiation using artificial neural networks, Renew. Energy **14**, 179 (1998)
9. S. Trajkovic, B. Todorovic, M. Stankovic, Estimation of FAO Penman C factor by RBF networks, Arch. Civil Eng. **2**, 185 (2001)
10. K.S. Reddy, M. Ranjan, Solar resource estimation using artificial neural networks and comparison with other correlation models, Energy Convers. Manage. **44**, 2519 (2003)
11. A. Angstrom, Solar and terrestrial radiation, Q. J. R. Meteorol. Soc. **50**, 121 (1924)
12. J.A. Prescott, Evaporation from a water surface in relation to solar radiation, Trans. R. Soc. South Aust. **64**, 114 (1940)
13. G.H. Hargreaves, Z.A. Samani, Estimating potential evapotranspiration, J. Irrig. Drain. Div. **108**, 225 (1982)
14. A.M. Tompkins, Impact of temperature and humidity variability on cloud cover assessed using aircraft data, Q. J. R. Meteorol. Soc. **129**, 2151 (2003)
15. K.K. Gopinathau, Solar sky radiation estimation techniques, Sol. Energy **49**, 9 (1992)
16. R.J. Stone, Improved statistical procedures for the evaluation of solar irradiation estimation models, Sol. Energy **51**, 289 (1993)

Cite this article as: Bilal Babar, Tobias Boström, Estimating solar irradiation in the Arctic, Renew. Energy Environ. Sustain. **1**, 34 (2016)

Paper II [redacted]

Solar Energy 170 (2018) 76–85

Contents lists available at ScienceDirect

Solar Energy

journal homepage: www.elsevier.com/locate/solener

Evaluating CM-SAF solar radiation CLARA-A1 and CLARA-A2 datasets in Scandinavia

Bilal Babar*, Rune Graversen, Tobias Boström

Energy and Climate group, Department of Physics and Technology, The Arctic University – University of Tromsø, Norway

ARTICLE INFO

Keywords:
CLARA A1 and A2
Scandinavia
ECMWF
Arctic
Solar radiation estimation
Polar orbiting satellites

ABSTRACT

Estimating/retrieving solar radiation through satellite-based remote sensing provides larger spatial coverage compared to other methods. Accurate estimates of incoming solar radiation is important when planning new solar energy installations. In addition, these estimates are also used in climate studies. Geostationary satellites are ideal for estimating solar radiation but cannot be used for high latitudes because of an unfavourable viewing angle; however, polar-orbiting satellites provide an alternative. **CL**oud, **A**lbedo **RA**diation edition 2 (CLARA-A2) is the latest retrieval product of cloud properties, surface albedo and surface solar radiation by Satellite Application Facility on Climate Monitoring (CM-SAF) based on Advance Very High Resolution Radiometer (AVHRR) observations from polar orbiting satellites. This data set covers the whole earth and provides daily and monthly averages. In this study, we have evaluated the CLARA-A2 data set and the previous version CLARA-A1 to *in-situ* high-quality observations from specific locations in Scandinavia, with a focus on solar radiation at high latitudes. The results show that both datasets perform within the target accuracies of CM-SAF, although the new data points, which were previously not available in CLARA-A1 due to snow-cover and cloud differentiation, have high deviations. Nevertheless, yearly average energy estimates are more accurate in CLARA-A2 because of these new points. For Swedish locations, mean absolute deviation (MAD) of 8.1 W m^{-2} and 8.7 W m^{-2} for CLARA-A1 and A2 respectively were calculated for updated values. Similarly, for Norwegian locations MAD of 8 W m^{-2} and 8.9 W m^{-2} were calculated for CLARA-A1 and A2. Overall, for all locations MAD lies at 8.1 W m^{-2} and 8.8 W m^{-2} for CLARA-A1 and A2, respectively. CLARA A2 has more temporal data points than CLARA A1, however, the MAD of the new data points that were not available in CLARA-A1 are 15.2 W m^{-2} and 17.7 W m^{-2} for Swedish and Norwegian sites, respectively.

1. Introduction

The surface radiation budget at the Earth plays a central role in climate monitoring and analysis of different meteorological parameters. Recent studies such as (Stroeve et al., 2014; Arndt and Nicolaus, 2014) make use of the surface radiation fluxes to indicate changing atmospheric and environmental conditions. In addition, surface radiation averages are used in the planning phase of the feasibility of solar energy conversion installations such as solar thermal or photovoltaic systems. Feasibility studies are important for choosing the optimal energy mix, as evident from the recent global status report by Renewable Energy Policy Network for the 21st Century (Ren21, 2017). The increase in the solar energy deployment in the past few years makes such datasets even more important for feasibility studies of future installations. In the Arctic regions there has been a growing interest in the use of clean and renewable energy sources, but the lack of reliable solar data hinders the socio-political decision-making processes. The focus of this paper is on validation and discussion of the improvements and shortcomings of the second edition of **CL**oud, **A**lbedo **RA**diation (CLARA) dataset for high latitude areas of Norway and Sweden. The retrieval quality of both data sets is tested against *in-situ* observations from locations at varying latitudes. In addition, these sites have different topography, especially in the Norwegian part.

Large solar power plants require preliminary data such as potential site locations and area-specific designs. The potential of a location is needed on a monthly and annual basis (Stoffel et al., 2010). The designs may vary, for example at high latitude locations, single or dual axis tracking increases the output energy by approximately 50% (Huld et al., 2010; Good et al., 2011). In addition, inter-annual variability of solar energy is used as a measure of change in received levels of radiation through a certain period to find uncertainties in the energy production at the locations where the solar energy units are planned (Kariuki and Sato, 2018). Long time series usually of the magnitude of multi-decadal order of solar radiation are analyzed in the preplanning

* Corresponding author.
E-mail address: bilal.babar@uit.no (B. Babar).

https://doi.org/10.1016/j.solener.2018.05.009
Received 29 November 2017; Received in revised form 23 March 2018; Accepted 2 May 2018
Available online 26 May 2018
0038-092X/

of power plants (Meyer et al., 2006). In most cases satellite-based databases or climate models are used to simulate solar-radiation parameters on a longer term, as these are usually not available from *in-situ* ground measuring stations. A common belief is that active solar energy production at high latitudes is not feasible since often the solar energy potential is underestimated. It is often neglected that the cold climate can be beneficial for solar energy harvesting as the efficiency of silicon solar cells increase at low temperatures (Skoplaki and Palyvos, 2009), and the presence of snow covers reflect solar radiation thereby boosting the output power. However, there are some challenges with solar energy at high latitudes such as a large seasonal variation in solar insolation, and a mismatch with the users demands. In this paper we focus on the challenge of accessing accurate solar irradiation data at high latitudes.

Various specialized databases are available for surface radiation estimation, including, European Solar Radiation Atlas (ESRA), solar data (SoDa), Satel-Light, Meteonorm, Photovoltaic Geographical Information System (PVGIS) etc (Dunlop et al., 2006). However, most datasets are based on geostationary satellites and therefore do not provide coverage above 60–65 degrees latitude. Others that use different satellite assimilation techniques take very few ground measuring stations into account, and thus cannot be considered as accurate for high latitudes. For locations above 60 degrees, retrieval methods based on observations from polar-orbiting satellites provide a solution, since these are shown to result in more accurate estimates than those obtained based on other remote sensing methods or empirical model estimation technique (Pinker and Laszlo, 1992; Besharat et al., 2013). As shown by Polo et al. (2016), satellite estimation of solar radiation has considerably improved and it is the second best option after the ground measurement methods. The Satellite Application Facility on Climate Monitoring (CM-SAF) provides multiple climate data records for cloud detection, albedo and surface radiation. CLARA data sets are one such product that can be used at high latitude locations because of its global coverage.

The most accurate *in-situ* instrument for recording global horizontal irradiance (GHI) is a pyranometer (Iqbal, 2012). In high-latitude Arctic regions, there are few meteorological stations and only a subset of these record solar radiation. The large distances between measurement hinder the exploitation of new sites for solar energy based on *in-situ* observations. Alternatively, solar radiation maps based on polar orbiting satellites can be used at these locations.

Some previous studies including Riihelä et al. (2015) and Urraca et al. (2017) have performed error statistics on the estimation of CLARA-A1 and CLARA-A2. In Riihelä et al. (2015), authors performed an extensive evaluation of CLARA-A1 and SARAH-A1 over Sweden and Finland, while in Urraca et al. (2017) a few sites from Norway were included. The novelty of this work lies in the comparison of the 2 datasets on Norway and Sweden over a larger number of sites and years. Moreover, the strength and weakness of the datasets are analyzed in depth.

This paper is organised as follows. Section 2 describes the sites used in the study and the sources of *in-situ* measurements. Section 3 describes methods used to process the data and the statistical evaluations performed. Section 4 presents the result and a discussion on these results. Section 5 concludes this work.

2. Sites

The locations used in this study are at different latitudes in Norway and Sweden. The reason for this is that the performance of Cloud, Albedo Radiation (CLARA) datasets can be assessed by taking into account that at higher latitudes there are more images provided by polar orbiting satellites (14 per day at poles). Coordinates of the locations, altitude and terrain information are provided in Table 1. The *in-situ* data used to validate both data sets are acquired from two different sources. For Norway, the data are from Norsk institutt for bioøkonomi (NIBIO), and for the Swedish locations, the data are from the database of Sveriges meteorologiska och hydrologiska institut (SMHI). Both databases contain average hourly measurement by Kipp and Zonen CPM11 or CMP13 pyranometers. The equipment is regularly maintained and datasets are quality controlled by the respective organizations. In case of SMHI, Baseline Surface Radiation Network (BSRN) routines by Long and Dutton (2010) are used for quality assurance. Missing or erroneous data are corrected by using meteorological variables described by Davies and McKay (1989). The network was upgraded in 2006–2007 and the average ratio between old and new measurements was found to be 0.997. More detail on the upgrade is given by Carlund (2011). NIBIO calibrates the equipment once every year and had a major overhaul in 2013. The equipment is inspected and maintained on daily or weekly basis (http://lmt.bioforsk.no/about). In this study, an additional quality check of the on-site observations was performed, and any data flagged for low quality were discarded. In addition, NIBIO measurements having more than 10% of hourly missing values in a year were discarded (see appendix for details about the years not included in the study).

Table 1
Information on the location, altitude and land cover type of the sites included in the study.

Norway	Latitude	Longitude	Altitude (m)	Land Cover Type
Tromsø	69.65	18.9	12	Island
Pasvik	69.45	30.04	27	Lakes/forest
Sortland	68.6	15.28	14	Coastal/fjords
Vågønes	67.28	14.45	26	Forest/Coastal
Tjøtta	65.83	12.43	10	Coastal/archipelago
Oslo	60.12	11.3	162	Rural/agricultural
Særheim	58.76	5.65	90	Inland/rural/agricultural
Lyngdal	58.13	7.04	4	Urban/Fjords/near coastal
Sweden	**Latitude**	**Longitude**	**Altitude (m)**	**Land Cover Type**
Kiruna	67.83	20.43	408	Sparse forest
Luleå	65.55	22.13	17	Coastal
Umeå	63.82	20.25	10	Near coastal
Borlange	60.48	15.43	140	Urban/forest
Stockholm	59.35	18.07	30	Coastal
Göteborg	57.70	12.00	5	Coastal
Lund	55.71	13.21	73	Urban

3. Method

3.1. Data source

CLARA edition 2 (CLARA-A2) by CM-SAF is the latest edition of CLARA datasets and was released in December 2016. The solar radiation estimates for CLARA are derived from the Advance Very High Resolution Radiometer (AVHRR) sensors on board METOP and NOAA polar orbiting satellites. The dataset is available for a 34 year period from 1st January 1982 to 31st December 2015, which is an extension of 6 years relative to the previous edition. The dataset covers the whole globe with a spatial resolution of 0.25 × 0.25 degrees on a regular lat-lon grid, which translates to 27.8 km at the equator. Average Surface Incoming Shortwave radiation (SIS) values are available for daily and monthly time resolutions. Instantaneous AVHRR images are processed to derive a spatio-temporal averaged dataset, consisting of cloud cover, surface albedo and surface-radiation products. The second edition is an improvement over the first edition because of the upgraded retrieval method and 6 years of additional data.

CLARA-A2 uses aerosol information, vertical integrated vapor and ozone, along with the surface albedo product to estimate incoming solar radiation (Jörg Trentmann, 2016). Estimation of surface albedo is a challenging task, which includes calculating top-of-the-atmosphere reflectance, classification of snow covered pixels, radiometric and

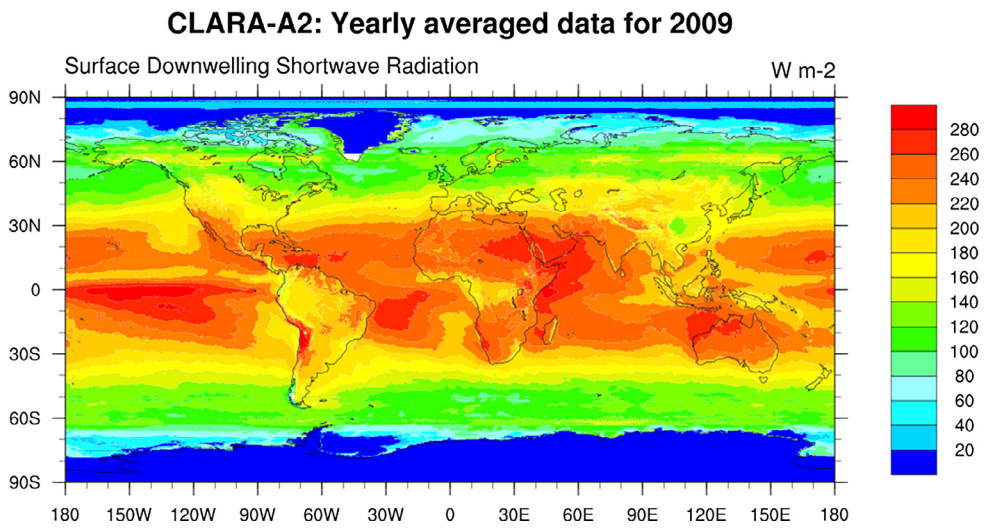


Fig. 1. CLARA-A2 yearly averaged solar irradiation data for 2009 on a horizontal surface.

geolocation topography correction, land use classification, etc. (Kati Anttila and Jääskelinen, 2016). In the case of high-latitude complex topography, a number of these methods are used to calculate the surface albedo including topography correction and classification of snow covered pixels. The viewing and illumination geometry at the satellite sensor becomes complex at low sun elevation. Such conditions increase the bidirectional surface reflectance thereby making the estimation process more complex (Kati Anttila and Jääskelinen, 2016). This aspect will be further discussed in later sections. Fig. 1 shows the CLARA-A2 yearly-averaged incoming solar radiation for 2009 on a horizontal surface.

Certain limitations exist in CLARA-A2; one of the main limitation is the availability of AVHRR observations. For calculating the daily averages, at least 20 observations are needed within a day and in each grid cell. In case of less than 20 images, the daily average field in question is filled with a value of $-999\ \mathrm{W\,m^{-2}}$ that represents a missing value. For a given grid cell, at least 20 days of observations is required to produce the monthly averages for SIS for a given grid cell. In case of availability of less than 20 days, the field is filled with a missing value.

A shortcoming of the dataset is the low number of satellites in the 1980s and the early 1990s, and for this reason only the period from 1995 and beyond is considered in this study. Another shortcoming includes the orbital drift of the satellites that results in different local observation times, which changes the observation conditions. Over Greenland the data quality was found to be insufficient to fulfil the threshold accuracy requirements, therefore, the southern tip of Greenland appears to be white which shows the area having missing values.

The major improvements in the latest CLARA edition on grid cell are from the cleaning and homogenizing of the basic level-1 AVHRR radiance data and the use of Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation (CALIPSO) Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP) cloud information. In the second edition, the cloud screening ability near poles is enhanced. Especially cloud detection over snow-covers is optimized and false cloud detection is reduced by using CALIOP cloud mask and CALIOP estimated cloud-optical thickness (Karlsson et al., 2017). A new dynamic aerosol optical depth (AOD) is used in CLARA-A2 surface albedo (SAL) calculations, which was previously set at a constant value of 0.1 (Kati Anttila and Jääskelinen, 2016). Moreover, the new edition uses wind speed in addition to sun zenith angles in SAL calculations (Kati Anttila and Jääskelinen, 2016). Digital elevation model used in this study is from NOAA (National Centers for Environmental Information). The snow depth data used to show the average snow depth of the areas in the analysis was obtained from ERA-Interim reanalysis (Dee et al., 2011).

3.2. Data processing

The ground-measured data used in this study are hourly averaged global horizontal irradiation. Refer to Section 2 for more details. The data from the SMHI database are quality controlled and flagged. From this dataset, sites flagged for bad quality were not used in the comparison. The NIBIO database is also quality controlled but not flagged. For Norway, hourly data for any year with large data gaps (10% or more of hourly values) were discarded. Missing values in this dataset were replaced by linear interpolation without taking diurnal solar elevation variation into account. For both NIBIO and SMHI, secondary standard pyranometers are used to record but these quality equipment have errors even when well-maintained and serviced. CMP11 Kipp and Zonen pyranometer have a flux measurement error of 2–5%. For monthly values lower uncertainty of 2% is expected in summer periods and 5% is expected in winter period (Wang et al., 2012). These uncertainties set an upper limit to the evaluation accuracy when estimates are compared with ground measured data (Riihelä et al., 2015).

Both CLARA datasets provide data of daily and monthly averages with a spatial resolution of 0.25 × 0.25 degrees (27.8 km × 27.8 km at the equator). Instead of fetching data for the closest grid point from the site locations, inverse distance weighted interpolation was used to calculated radiation values at precisely the site locations. Whenever the surrounding four grid points have more than 1 missing value for a certain time; the interpolation was replaced by a missing value of $-999\ \mathrm{W\,m^{-2}}$. By using this method, a slight improvement was observed in the overall deviations.

3.3. Statistical evaluation of estimations

Different statistical measures are used to evaluate the model deviations. The most widely used measure is the Root Mean Squared Deviation (RMSD). As an additional measure the BIAS or mean bias deviation (MBD) is used in the evaluation. Using MBD gives an insight in the general trends of under or over estimations. Mean absolute deviation (MAD) is also used for the evaluations of datasets. Because of the absolute values used in this measure, the negative and positive deviations do not cancel out each other as in the MBD. This is a good measure to compare different models as the one with smaller MAD will be the more reliable for estimations (Last et al., 2001).

Table 2
CLARA-A1 and CLARA-A2 monthly averaged comparison results from 1995 to 2009. The deviations are represented by root mean square deviation (RMSD), mean bias deviation (MBD) and mean absolute deviation (MAD). Numbers in parenthesis are the results for daily mean values. The table shows the results for Norway and Sweden seperately along with results from all sites.

	RMSD ($W m^{-2}$)		MBD ($W m^{-2}$)		MAD ($W m^{-2}$)	
Location	A1	A2	A1	A2	A1	A2
			Norwegian Locations			
Tromsø	18	16	3.4	−4	4.2	8.7
	(46)	(24)	(4.3)	(−3)	(10.4)	(12)
Pasvik	11	16	1	−2.9	3.3	6.2
	(36)	(22)	(2.1)	(−2)	(8.6)	(8.8)
Sortland	11	18	−3.7	−11.3	4.4	11.5
	(21)	(24)	(−2.8)	(−10.7)	(7.6)	(14.3)
Vågønes	13	11	1.3	−2	4.3	5.4
	(35)	(17)	(2.8)	(−1)	(9.9)	(9.6)
Tjøtta	8	7	2.2	−1.3	3.7	4.2
	(33)	(16)	(3.6)	(−0.3)	(10.7)	(8.4)
Oslo	9	10	−2.3	−3.7	4.1	5.8
	(33)	(18)	(−0.6)	(−2.3)	(12.5)	(10.4)
Særheim	7	7	1.2	−1.9	4.3	4.4
	(31)	(16)	(2.7)	(−0.3)	(13.8)	(9.5)
Lyngdal	12	20	−2.7	−7.6	6.4	9.5
	(24)	(34)	(−1.7)	(−6.6)	(11.6)	(13.9)
All Norwegian locations	11	14.2	−0.1	−5.6	8	8.9
	(34)	(24.9)	(1.9)	(−4.1)	(18.7)	(13.5)
			Swedish Locations			
Kiruna	8	18	−0.5	−0.5	2.6	7.8
	(29)	(24)	(0.8)	(0.8)	(7.5)	(11.1)
Luleå	9	9	1.2	−0.8	3.5	4.3
	(27)	(16)	(2.7)	(0.1)	(8.4)	(7.4)
Umeå	8	11	0.5	−4	3.7	6.8
	(27)	(17)	(2.5)	(−2.6)	(8.9)	(9.3)
Borlange	9	9	−1	−3.6	4	9.4
	(27)	(17)	(0.7)	(−2.1)	(10.8)	(5.5)
Stockholm	8	9	2.4	3.2	4.7	5.5
	(28)	(18)	(4.6)	(1.7)	(12.5)	(9.9)
Göteborg	7	7	1.9	0.8	4.5	4.5
	7 (25)	(16)	(3.6)	(2.3)	(12.4)	(9.4)
Lund	9	8	−2.1	−1.8	4.9	5.2
	(25)	(17)	(−0.9)	(−0.1)	(11)	(10.4)
All Swedish Locations	11.7	13	0.5	−2.5	8.1	8.7
	(41.6)	(46.5)	(2.9)	(−1.1)	(21)	(17.4)
All Locations	**11.4**	**13.5**	**0.2**	**−3.8**	**8**	**8.8**
	(38.7)	**(38.5)**	**(2.5)**	**(−2.4)**	**(20)**	**(15.6)**

4. Results and discussion

Table 2 shows the results of the statistical evaluation performed over the period of 1995–2009 over Sweden and Norway. The evaluations are arranged in decreasing latitudes in the tables. For most of the sites, CLARA-A2 provides lower RMSD values for daily means, but for monthly means, CLARA-A1 performs better or very similar to CLARA-A2.

In terms of biases, CLARA-A1 performs better at most of the sites. At some locations though the opposite pattern is found, but overall the Swedish locations show an overestimation and the Norwegian locations an underestimation. In a previous work by Riihelä et al. (2015), a similar overestimation was reported for CLARA-A1 in Sweden. The frequency of observations of the satellite also contributes to the errors, where 20 images are used to estimate daily and monthly averages, while the available frequency of ground observations is once every hour.

For both data sets, the threshold, target and the optimal accuracy is 15, 10 and 8 $W m^{-2}$ respectively, for monthly averages and 30, 25 and 20 $W m^{-2}$ for daily averages as described in Karlsson et al. (2012), Jörg Trentmann (2012), Karlsson et al. (2017), and Jörg Trentmann (2016), respectively. The MAD in Table 2 indicates that all the results are well within these specified thresholds, and most of the sites show an optimal accuracy of 8 and 20 $W m^{-2}$ for monthly and daily averages, respectively. For Norwegian locations, monthly MAD of 8 $W m^{-2}$ was recorded for CLARA-A1 while for CLARA-A2 it was 8.9 $W m^{-2}$ and for Swedish locations, monthly MAD was 8.1 $W m^{-2}$ for CLARA-A1 and 8.7 $W m^{-2}$ for CLARA-A2 $W m^{-2}$. The overall MAD for CLARA-A1 and A2 for daily averages were 20.05 $W m^{-2}$ and 15.65 $W m^{-2}$ and for monthly averages 8.06 $W m^{-2}$ and 8.82 $W m^{-2}$, which is also within the limits of CM-SAF. For most of the sites the daily accuracies are improved in the later CLARA edition relative to the former, while CLARA-A1 performs better on monthly accuracies for most of the sites. Furthermore, CLARA-A2 has more monthly and daily mean data points than CLARA-A1, especially at higher latitudes as shown by the Hovmöller diagram in Fig. 2. Higher latitudes have more snow covers, which are estimated more frequently in CLARA-A2. The availability of the datasets will be elaborated further in the subsequent sections.

Polar orbiting satellites follow a sun synchronous orbit in which the temporal resolution of sensing increases with latitude. About 14 daily observations are recorded close to the poles per satellite swath, whereas only two observations are available close to the equator (Karlsson et al., 2017). At latitudes below 65 degrees the number of images captured by polar orbiting satellites is not high enough to obtain the daily means when the day length is short, while the availability rises again above 65 degrees because of the overlapping of the satellite swath. At even higher latitude, the coverage is larger but the main challenge at such

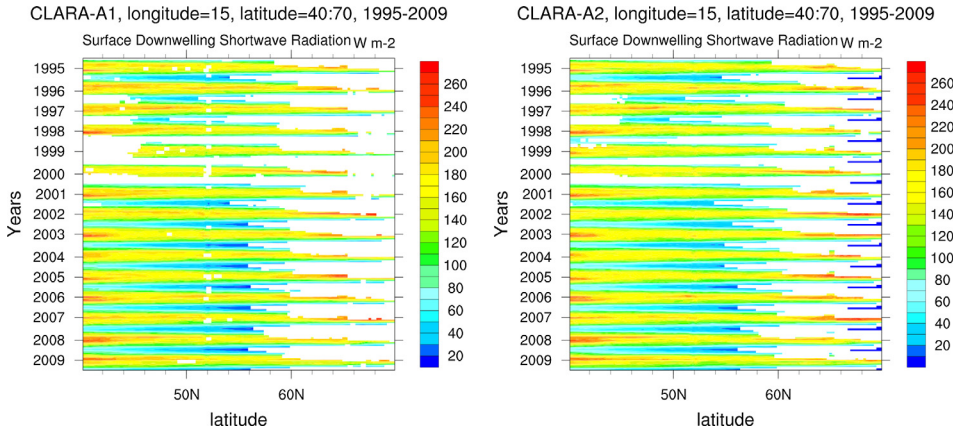


Fig. 2. Hovmöller plots for CLARA-A1 and A2 datasets for the included years in the study. The plots are centered at 10 degrees longitude and span from 40 to 70 degrees latitude.

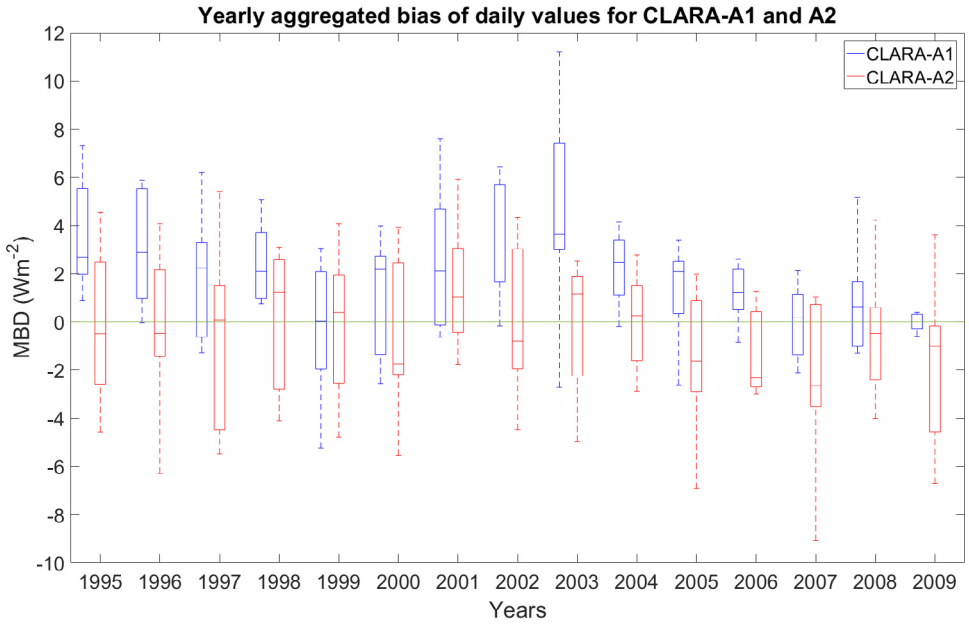


Fig. 3. Box plot showing the inter annual stability of CLARA-A1 and A2. The stability is shown in terms of mean bias deviation. 25th and 75th percentile values are shown by the length of the box.

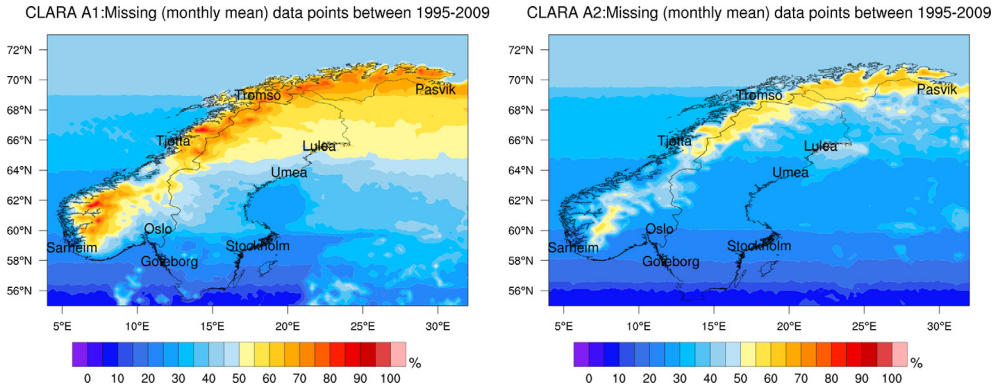


Fig. 4. Percentage of monthly averaged data missing values in the datasets. Figure on the left shows the missing points in CLARA-A1 dataset between 1995 and 2009. Figure on the right shows the missing points in CLARA-A2 for the same period.

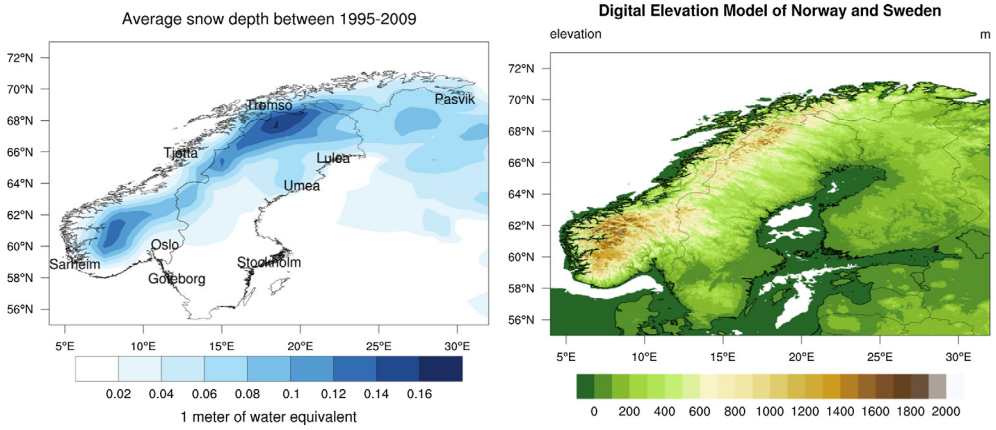


Fig. 5. Average snow depth between 1995 to 2009 from ERA-Interim and topography. Larger snow depth occurs at complex terrains, and most missing data points lie in such regions.

high latitudes is the snow covered surfaces (Urraca et al., 2017). In this study, the Norwegian locations have snow covers in addition to a very complex terrain including a high number of fjords and mountains (see Fig. 5). It is highly likely that satellite retrieval estimation methods deteriorate on mountain regions because the spatial resolution of incident light on satellite sensor is not high enough to compensate for the complex terrain, while sudden changes in weather conditions due to mountains are not compensated for with low sensing frequency as in

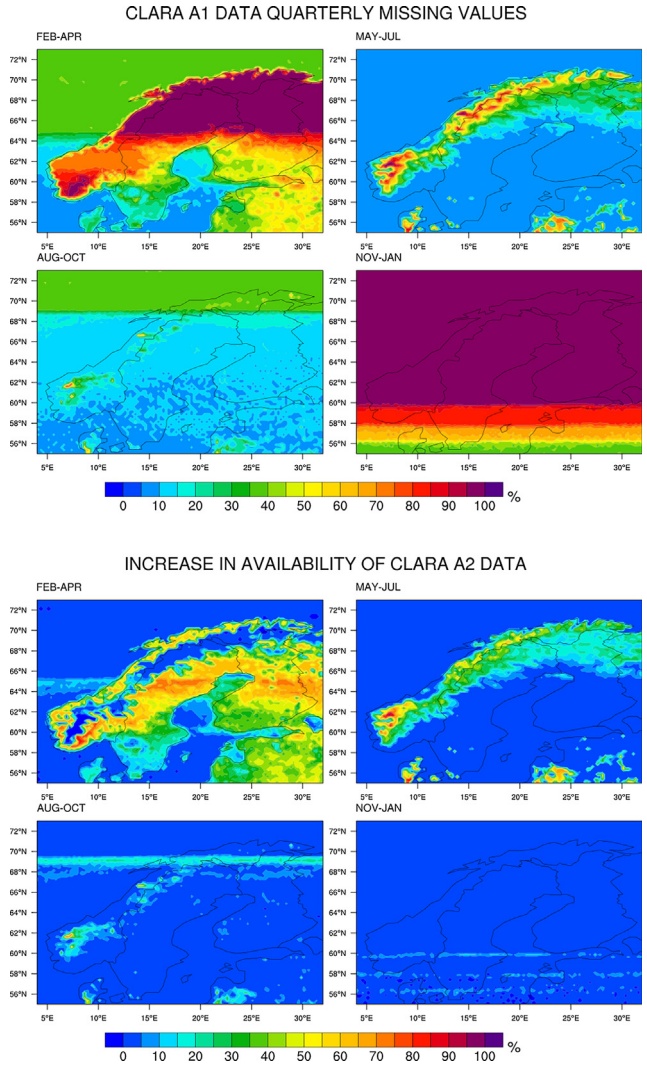


Fig. 6. The top figure shows the percentage of monthly missing data in CLARA-A1 in each quarter. The lower figure shows the percentage increase in the availability of CLARA-A2 dataset in each quarter. The highest increase is in the areas that have complex topography in addition to snow covers.

the case of polar orbiting satellites.

This study is conducted on mountainous regions with snow covers, which not only introduces random errors but also negative biases. Furthermore, because the satellite estimation methods use the visible spectrum channels for the detection of clouds, the sensors cannot differentiate between clouds and snow cover, which further contributes to increasing the errors (Urraca et al., 2017). However, 0.6 and 0.8 μm channels are used separately in order to detect snow covers and calculating the albedo (Kati Anttila and Jääskelinen, 2016). Albedos for snow are high in the near ultra-violet and visible spectrum, but it starts dropping drastically in the near infra-red region between 0.8 and 1.5 μm (Wiscombe and Warren, 1980). Most of the high latitude sites in this study have snow cover for a large part of the year. Which implies a further increase of errors in the datasets. Although the new dataset have more coverage over snow-covers, which was previously not available in CLARA-A1, but such new values have large errors. These large errors are likely due to the differentiation between snow and cloud covers (see Fig. 5).

4.1. Inter-annual stability

As discussed earlier, inter-annual stability of a dataset provides insight into the uncertainties associated with the energy production of solar energy plants. Areas where typical ground measuring equipment are not available can take advantage of datasets provided by CM-SAF. Therefore, such datasets should be consistent throughout the periods of investigation. In Fig. 3 the box plot of MBD of both CLARA-A1 and A2 datasets are shown. It can be seen from the figure that the CLARA-A2 dataset has lower median bias than the CLARA-A1 dataset, with median values being closer to the zero bias. The CLARA-A2 dataset has more extreme minimum values, compared to CLARA-A1, while the maximum values are in most cases better in the CLARA-A2 dataset. Moreover, the 25th and 75th percentile values in CLARA-A2 data set lies approximately around -2 and $2\,\mathrm{W\,m^{-2}}$, while in CLARA-A1 these values are approximately around 0 and $4\,\mathrm{W\,m^{-2}}$. These results show that the newer edition of CLARA has more stability in terms of biases over the years included in the study period.

4.2. Data availability

One of the improvements of CLARA-A2 is the differentiation of snow-covered surfaces from cloud covers in the surface albedo calculations. Both CLARA datasets do not provide coverage over snow-covered surfaces (Riihelä et al., 2015; Karlsson et al., 2017) and such time periods are filled with missing values. Nevertheless, because of the improvement in surface albedo calculations, CLARA-A2 provide more data points than CLARA-A1. The additional data points in CLARA-A2 are mostly from the snow-cover time periods, hence there is not much improvement in the overall skills. In most cases, there is a higher degree of deviation at such locations, which further increase the deviations as a whole. As shown in Fig. 4, CLARA-A1 has roughly between 50 and 80% missing values in Norway and around 40 to 60% missing values in Sweden. In comparison CLARA-A2 has approximately 30 to 60% missing data in Norway and 20 to 50% missing data in Sweden. This further explains the results in Table 2, where CLARA-A1 performs better than CLARA-A2 and that the skills for the Swedish locations are better than those at the Norwegian locations. The complex topography of Norwegian locations along with a high percentage of snow covers at these areas have resulted in inaccurate estimations that previously were replaced by missing values and thus not taken into account in statistical evaluations. Fig. 5 below shows the average snow depth in the study period between 1995 and 2009 along with a digital elevation model of the study area. By comparing Fig. 5 with the maps in Fig. 4, it can be seen that in CLARA-A1 snow-covers correspond to missing values.

Similarly, in CLARA-A2 there are less missing values on snow covered grid points, but still the highest amount of missing data are found on the higher snow-depth grid points and high elevation locations.

4.3. Seasonal variations in the datasets

To further investigate the datasets, seasonal variation of both datasets were calculated. Data from 1995 to 2009 were divided into quarterly datasets by assigning the months from February to April to the 1st quarter, May to July to the 2nd quarter, August to October to the 3rd quarter and November to January to the 4th quarter. In this manner, we could separate the darker and snow covered periods from the summer months.

Fig. 6 illustrate the quarterly frequency of missing data in the CLARA-A1 data set and illustrates the increase in the availability of data points in the new edition compared with the previous edition. It further illustrates that due to the fact that most of the northern parts of Norway and Sweden has snow-covers, most of the missing data point in CLARA-A2 lie in these regions. The availability has increased in these northern location in CLARA-A2 when compared to CLARA-A1, though not so much in the high snow-depth mountain regions (see Fig. 5). The highest amount of missing values lie in the February to April months when the polar night has ended and the snow is melting.

Table 3 gives the seasonal deviations of the two datasets. It can be

Table 3
Quarterly deviations for CLARA-A1 and CLARA-A2 datasets. The table shows the seasonal variation in the biases of both datasets. Monthly average values for the years included in the study were divided into four quarters that are denoted by Q. CLARA A1 and A2 datasets are denoted by A1 and A2, respectively.

Norway/Quarter		RMSD ($W m^{-2}$)		MBD ($W m^{-2}$)		MAD ($W m^{-2}$)		Correlation	
		A1	A2	A1	A2	A1	A2	A1	A2
Tromsø	Q1	–	11.6	–	−5	–	5	–	0.9
	Q2	26	25.2	5.6	−11.3	5.9	19.8	0.60	0.9
	Q3	14.6	11.4	7.8	1.2	11	8.7	0.98	0.9
	Q4	–	2.3	–	−1.2	–	1.2	–	–
Pasvik	Q1	–	3.2	–	−0.6	–	0.6	–	–
	Q2	13.9	27.5	4.3	−5.9	4.8	13	0.95	0.7
	Q3	10.2	11.5	0	−3.5	8.4	9.7	0.98	0.9
	Q4	–	3.7	–	−1.4	–	1.4	–	–
Sortland	Q1	22.3	22.9	−1.6	−11.6	1.6	11.6	–	0.97
	Q2	12	22.6	−5.1	−17.9	6.7	18.4	0.9	0.97
	Q3	10	13.7	−7.6	−11.9	8.5	12	0.9	0.99
	Q4	4.7	9.6	−0.5	−3.7	0.6	4	0.9	0.88
Vågønes	Q1	5.1	8	0.2	−4	0.2	4	–	0.99
	Q2	12	8.4	6.9	3	6.9	6.8	0.9	0.99
	Q3	15	15.5	−1.4	−5.3	8.6	8.8	0.9	0.96
	Q4	9.4	7.2	−0.4	−2	1.4	2.1	0.6	0.94
Tjøtta	Q1	9.4	7.8	0.6	−1.9	0.6	2.2	–	0.89
	Q2	10.8	7.7	6.9	2.1	8.5	6.6	0.9	0.98
	Q3	6.5	7.8	1	−3.9	5	6.4	0.9	0.99
	Q4	3	3.4	0.3	−1.5	0.6	1.5	0.9	0.99
Oslo	Q1	10.7	29.4	−1.2	−12.5	1.2	12.5	0.97	0.87
	Q2	21.5	20.6	−6	−3.4	10	8.3	0.88	0.89
	Q3	12.4	11.7	−5.3	−3.8	9.1	8.5	0.97	0.97
	Q4	6.1	9.3	−1	−2.5	1.5	2.5	0.93	0.93
Særheim	Q1	5.7	6.7	1.5	−3	2.9	3.3	0.99	0.98
	Q2	6.8	5.8	3.3	1.7	5.6	4.5	0.99	0.99
	Q3	7.9	9	−0.3	−3.6	7	7.3	0.99	0.99
	Q4	3.7	5.9	0.4	−2.5	1.5	2.6	0.98	0.99
Lyngdal	Q1	10.2	34.5	−0.5	−10.8	2.9	10.8	0.97	0.66
	Q2	12.5	13.8	−1.2	−4.4	9.7	10.9	0.96	0.96
	Q3	14.4	16.6	−8.2	−11.2	10.2	11.9	0.97	0.98
	Q4	8.3	11.5	−1.1	−4.2	2.8	4.2	0.90	0.90
Kiruna	Q1	–	15	–	−4.3	–	4.3	–	1.00
	Q2	8.3	29.6	0.9	−12.6	3.1	18.3	0.94	0.84
	Q3	8	8.7	−3.1	−3.3	6.4	6.8	0.99	0.99
	Q4	3.1	4	0.2	−1.7	0.7	1.7	0.86	0.99
Luleå	Q1	–	–	–	–	–	–	–	–
	Q2	12.5	12.1	6	2.6	7	7.6	0.94	0.96
	Q3	7.9	8.8	−1.3	−3.5	6.3	7.3	0.99	0.99
	Q4	3.7	4.5	−0.1	−2.3	0.9	2.2	0.76	0.99
Umeå	Q1	2.3	13.9	0.3	−6.4	0.4	6.4	0.84	0.99
	Q2	9.3	11.6	4.8	−1.5	6	9.3	0.97	0.98
	Q3	9.3	10	−2.7	−5.2	7.5	8.6	0.99	0.99
	Q4	3.3	5.4	−0.3	−3	0.8	3	0.85	0.99
Borlange	Q1	4.2	11	−0.8	−6.5	0.9	6.5	1.00	0.99
	Q2	7.7	6.4	0.2	−1	6	5.1	0.98	0.99
	Q3	9.7	9.6	−4.3	−5.6	7.3	8.2	0.98	0.99
	Q4	10	9.7	0.8	−1.2	2	2.3	0.61	0.68
Stockholm	Q1	14.1	13.5	1.2	−2	5.2	6.1	0.88	0.90
	Q2	22	23	5	7	18.2	19.3	0.81	0.82
	Q3	29.9	31.4	−1.2	0.2	22.9	23.2	0.81	0.80
	Q4	10	11.1	0.6	−2	3.9	4.2	0.84	0.83
Göteborg	Q1	5.5	3.5	2.5	−0.9	3	1.8	1.00	1.00
	Q2	9	8.9	5.8	6.4	7.5	7.7	0.99	0.99
	Q3	7.5	7.6	−0.9	0.01	6.1	6.2	0.99	0.99
	Q4	4	6.2	0.2	−2.4	1.5	2.5	0.98	0.98

(*continued on next page*)

Table 3 (*continued*)

Norway/Quarter		RMSD (W m^{-2})		MBD (W m^{-2})		MAD (W m^{-2})		Correlation	
		A1	A2	A1	A2	A1	A2	A1	A2
Lund	Q1	5	6.6	1.8	−2.8	2.7	4	1.00	0.99
	Q2	8.5	7.5	−1.5	2.9	4.9	6.2	0.98	0.99
	Q3	12.6	9.1	−8.8	−4	9.4	7.1	0.98	0.99
	Q4	4.6	6	−0.03	−3.3	2.5	3.4	0.99	0.99

Table 4
Analysis of the new and updated solar radiation values in CLARA-A2 for Norwegian and Swedish locations. The column marked with New are the values which were not available in CLARA-A1 (shown in last column, No. of new values), while the updated values are the ones which were available in CLARA-A1 but were updated in CLARA-A2.

Norwegian Location	RMSD (W m^{-2})		MBD (W m^{-2})		MAD (W m^{-2})		No. of new values
	New	Update	New	Update	New	Update	
Tromsø	25	14	−4.7	0.9	5	3.2	20
Pasvik	44	12	−2.3	−0.1	2.3	3.4	8
Sortland	30	15	−4.4	−6	4.5	6.1	18
Vågønes	9	13	−1.1	−0.7	1.3	3.9	23
Tjøtta	7	8	−0.7	−0.5	0.7	3.4	14
Oslo	16	8	−2	−1.7	2	3.8	15
Særheim	9	7	−0.3	−1.6	0.3	4.2	4
Lyngdal	46	14	−2	−5.6	2	7.4	12
ALL SITES	25.9	11.5	−17	−3.7	17.7	8.3	114 (12%)
CLARA-A1 (All Included)	11		−0.1		8		
CLARA-A2 (All included)	14.2		−5.6		8.9		
Swedish Location	RMSD (W m^{-2})		MBD (W m^{-2})		MAD (W m^{-2})		No. of new values
	New	Update	New	Update	New	Update	
Kiruna	37	8	−4.7	−0.6	4.7	2.9	26
Luleå	17	9	−0.5	−0.1	0.5	3.6	6
Umeå	18	8	−2.4	−1.3	2.4	4.1	27
Borlange	12	8	−1.1	−2.4	1.1	4.4	18
Stockholm	6	23	−0.2	2	0.2	13	8
Göteborg	6	7	−0.1	0.9	0.1	4.4	6
Lund	9	7	−0.6	−1.1	1.1	4.1	30
ALL SITES	20.6	11.9	−14.5	−0.7	15.2	8.3	121 (9.6%)
CLARA-A1 (All included)	11.7		0.5		8.1		
CLARA-A2 (All included)	13		−2.5		8.7		

seen that in the 1st and 2nd quarter, CLARA-A2 provides more valid data points than does CLARA-A1 (see also Fig. 6). Missing data or no valid value at grid points means that these months are not taken into account when making any of the calculations in the study. When compared to the snow-depth map in Fig. 5, the regions of missing values lie approximately on the areas having higher snow-depth and complex topography. The 1st and 4th quarters have special conditions, where the 1st quarter has low sun-elevation angles and the 4th quarter includes the polar-night period. Moreover, the 1st and 3rd quarter have similar and opposite sun elevation angles (in the 1st quarter the solar elevation increases while in the 3rd quarter it decreases) but the 1st quarter has more snow-cover than the 3rd quarter. It also shows that in the 1st quarter both the MBD and MAD are larger in CLARA-A2 than CLARA-A1. Low RMSD values are observed below 60 degrees in Swedish locations but not in Norwegian locations. The MBD or bias is mostly negative for CLARA-A2, with high values for Norway than for Sweden. However, due to the unavailability of data in some high latitude locations it was not possible to calculate the deviations. In the 2nd quarter, CLARA-A1 has better RMSD measures until around north of 60 degrees after which CLARA-A2 either starts improving or provides similar values as CLARA-A1 (except for Pasvik, Sortland and Kiruna). Similarly, CLARA-A1 again provides better MBD and MAD values. In the 3rd and 4th quarters, all the measures are either similar in both the datasets or slightly worse in CLARA-A2 for both Norwegian and Swedish location. Based on the observations it can be said that although CLARA-A2 has more coverage over snow-covered areas it still provides large deviations at high latitude locations.

4.4. Analysis of the new and updated monthly average values in CLARA-A2

By comparing CLARA-A1 and A2, it can be seen that there are two major changes in the availability of data. First, there are fewer missing values in A2 and secondly, the adjacent grid point values are also updated in CLARA-A2 due to the use of different methods of estimation. This section provides an evaluation of the new and updated monthly means estimations separately. The values marked with “New” are the values which were not available in CLARA-A1 (marked as a missing values) but that are available in CLARA-A2. The values marked with “Updated” are those values which were available in CLARA-A1 but these got updated because of the use of new algorithms. In this way we could separately analyse the improvement of CLARA-A2. Table 4 shows the RMSD, MBD, MAD and the number of new values in CLARA-A2. For the newly added added data points in CLARA-A2 the MAD target accuracies for all locations are above the limits (17.7 W m^{-2} for Norway and 15.2 W m^{-2} for Sweden). Individually for both Sweden and Norway, the updated values are very similar and within the target (8.3 W m^{-2} for both Norway and Sweden). Table 4 also shows the overall accuracies of both datasets for all Norwegian and Swedish locations. Overall accuracies for both datasets also are within the limits.

Furthermore, the new values in CLARA-A2 have a constant negative bias that shows the underestimation in these values. The cause for this underestimation can be attributed to the inaccurate detection of snow-covers. The RMSD section of the table shows that the new values have very high deviations for high-latitude locations in both countries; nevertheless, the updated values have relatively low RMSD because of the upgraded retrieval method and absence of snow-covers.

4.5. Analysis of annual energy estimates

The total annual energy estimate at a site is an important parameter for planning purposes. In addition to daily and monthly averages that are used in the inter-annual stability for energy production, annual energy averages give an insight into the total energy that can be harvested at potential site locations. Table 5 shows the RMSD, MBD and MAD of yearly averaged hourly solar irradiances of CLARA-A1 and A2. In this analysis, CLARA-A2 performs considerably better than CLARA-A1 in all areas. Moreover, average annual energy is also listed for both CLARA datasets and *in-situ* values. For calculating yearly energy values, mean hourly values from ground-measured data and mean daily values from CLARA datasets were used. By comparing the energy potential estimates it can be seen that CLARA-A2 provides better estimates than CLARA-A1. The energy estimates are better in CLARA-A2 due to the fact

Table 5
This table shows annual average solar radiations error analysis for CLARA-A1 and A2 for Norwegian and Swedish locations in terms of RMSD, MBD and MAD. The portion of the table labelled as Power is expressed in W m^{-2}. The right side of the table shows the annual average energy estimates of CLARA-A1, A2 and ground-observed data expressed in kW h m^{-2} y.

Norwegian Locations	Power						Energy		
	RMSD (W m^{-2})		MBD (W m^{-2})		MAD (Wm^{-2})		A1 (avg) kW h m^{-2} y	A2 (avg) kW h m^{-2} y	Obs (avg) kW h m^{-2} y
	A1	A2	A1	A2	A1	A2			
Tromsø	69.7	9.8	68.7	7.7	68.7	7.7	469.1	643.7	687.4
Pasvik	65.7	12	65.3	9.3	65.3	9.8	497.4	544.6	718.2
Sortland	50.7	4.7	48.8	2.1	48.8	3.7	600.3	664.8	780.4
Vågønes	53.7	13.8	53.1	12.9	53.1	12.9	600	724	733.9
Tjøtta	57.5	25.6	56.8	25.3	56.8	25.3	698.9	749.9	768.2
Oslo	48.2	31.3	47.2	30.2	47.2	30.2	827.5	902.4	948.7
Særheim	29.7	21.6	28.8	21.3	28.8	21.3	913.8	901.7	921.7
Lyngdal	31.9	21.7	29.7	17.2	29.7	18.7	915.7	939.8	1032.9
Swedish Locations	**Power**						**Energy**		
	RMSD (W m^{-2})		MBD (W m^{-2})		MAD (Wm^{-2})		A1 (avg) kW h m^{-2} y	A2 (avg) kW h m^{-2} y	Obs (avg) kW h m^{-2} y
	A1	A2	A1	A2	A1	A2			
Kiruna	48.6	9.1	47.5	8.4	47.5	8.4	525	654.7	804.5
Luleå	62.3	34.5	61.5	34.3	61.5	34.3	704.3	728.1	895.8
Umeå	51	18.9	48.6	17.4	48.6	17.4	777.2	860.4	916.7
Borlange	43.7	29.8	42.7	28.9	42.7	28.9	846.7	893.3	937.2
Stockholm	38.3	32.8	36.6	30.4	36.6	30.4	984.5	998	993.4
Göteborg	32.1	26.9	30.3	24.9	30.3	24.9	968.3	966.5	969.6
Lund	18.8	17.4	4.9	9	13.4	11.8	791.1	1013	1034.7

that it provides more data points than CLARA-A1. Fewer data points in the time series means that the energy estimates for CLARA-A1 results in lower estimates than both CLARA-A2 and ground observed data.

The energy estimates provided in Table 5 are for the yearly solar radiation received on a horizontal plane per area averaged over the study period. At high latitude locations, the elevations of the sun are often very low and consequently the horizontal solar density decreases. The difference between high and low latitude locations is considerably less when looking at an optimally inclined or a tracking surface.

5. Conclusion

In this work, we evaluated two datasets derived from polar orbiting satellites. CLARA-A2, the newer version of the CM-SAF polar orbiting satellite-based database, is derived with a procedure including improvements in cloud cover and snow cover distinction; hence, there are more data points taken into account in the new dataset. Still, missing values exist in the new dataset due to lack of differentiation between clouds and snow covers. However, the newer edition does not considerably improve the estimates for Northern Scandinavia. The evaluation metrics used in the study provides an insight into the performance of these datasets. CLARA-A2 is observed to provide underestimation at most locations, while CLARA-A1 provides more positive biases. This underestimation can be associated with the snow and cloud detection and the difficulties to differentiate between the two, which hopefully will be further improved in CLARA-A3, the next edition of this dataset that is planned to be launched in 2020. The CLARA-A2 dataset has less intra-annual variability than CLARA-A1, and along with the spatiotemporal resolution, it provides a more reliable dataset for areas below 60 degrees latitude. For the magnitude of errors presented in this study, consideration should be given to the complex topography especially in the case of Norwegian sites. Table 2 shows that MBD and MAD values are predominantly higher at Norwegian location. However, at most locations the target monthly average accuracies of 9 W m^{-2} for CLARA-A2 and 10 W m^{-2} for CLARA-A1 are achieved, along with daily average accuracies of 18 W m^{-2} for CLARA-A2 and 20 W m^{-2} for CLARA-A1. A quarterly deviation analysis shows that due to the complex topography and snow cover in Norwegian locations, CLARA-A2 does not provide more accurate estimates than CLARA-A1. Analysis on the new data points of CLARA-A2, that were previously not available, shows that these new values have very high deviations. Nevertheless, yearly energy estimates of CLARA-A1 are predominantly lower than CLARA-A2 estimates since there are simply more data points in CLARA-A2. To conclude, even if CLARA-A2 has a higher negative bias than CLARA-A1 at the specific common data points, CLARA-A2 still has more accurate yearly energy estimates because it has more data points than CLARA-A1.

Acknowledgements

The authors would like to thank Jörg Trentmann and CM-SAF for providing support and assistance regarding the datasets. We would also like to thank SMHI and NIBIO for providing ground measured solar radiation data. This work is supported by Troms county and industry development fund under the project title, “Renewable energy in the arctic – academy and business in a joint effort” RDA12/46.

Appendix A

Years within the studying period of 1995 to 2009 not included in this work (see Table 6).

Table 6
Detail of years not included in the study for each location.

Kiruna	N.A
Luleå	N.A
Umeå	N.A
Borlange	N.A
Stockholm	1998
Göteborg	N.A
Lund	N.A
Tromsø	1995, 1996, 2000, 2001, 2002, 2006, 2007, 2008
Pasvik	1995, 1996, 2006, 2007
Sortland	1995, 1996, 1997, 2000, 2003, 2007
Vågønes	1995, 1996, 1997, 2007
Tjøtta	1995, 1996, 1997, 2006, 2007
Oslo	1995, 1996, 1997, 1998, 2006, 2007
Særheim	1995, 1996, 2000, 2006, 2007
Lyngdal	1995, 1996, 2003

References

Arndt, S., Nicolaus, M., 2014. Cryosphere 8, 2219–2233.
Besharat, F., Dehghan, A.A., Faghih, A.R., 2013. Renew. Sustain. Energy Rev. 21, 798–821.
Carlund, T., 2011. Upgrade of SMHI's Meteorological Radiation network 2006-2007: Effects on Direct and Global Solar Radiation. SMHI.
Davies, J., McKay, D., 1989. Sol. Energy 43, 153–168.
Dee, D.P., Uppala, S., Simmons, A., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., Bauer, P., et al., 2011. Quart. J. Roy. Meteorol. Soc. 137, 553–597.
Dunlop, E.D., Wald, L., Suri, M., 2006. Solar Energy Resource Management for Electricity Generation from Local Level to Global Scale. Nova Science Publishers Inc.
Good, C., Persson, H., Kleven, Ø., Norton, M., Boström, T., 2011. In: Presentert på 26th European Photovoltaic Solar Energy Conference and Exhibition.
Huld, T., Cebecauer, T., Šúri, M., Dunlop, E.D., 2010. Prog. Photovoltaics Res. Appl. 18, 183–194.
Iqbal, M., 2012. An Introduction to Solar Radiation. Elsevier.
Jörg Trentmann, R.M., 2012. C. Team.
Jörg Trentmann, S.K., 2016. C. Team.
Kariuki, B.W., Sato, T., 2018. Renew. Energy 116, 88–96.
Karlsson, K., Riihelä, A., Müller, R., Meirink, J., Sedlar, J., Stengel, M., Lockhoff, M., Trentmann, J., Kaspar, F., Hollmann, R., et al., 2012. Satell. Appl. Facility Clim. Monit. 1. http://dx.doi.org/10.5676/EUM_SAF_CM/CLARA_AVHRR.
Karlsson, K.-G., Anttila, K., Trentmann, J., Stengel, M., Meirink, J.F., Devasthale, A., Hanschmann, T., Kothe, S., Jääskeläinen, E., Sedlar, J., et al., 2017. Atmos. Chem. Phys. 17, 5809–5828.
Kati Anttila, A.R.T.M., Jääskelinen, Emmihenna, 2016. C. Team.
Last, J.M., Abramson, J.H., Freidman, G.D., 2001. A Dictionary of Epidemiology, vol. 4 Oxford University Press, New York.
Long, C.N., Dutton, E.G., 2010.
Meyer, R., Lohmann, S., Schillings, C., Hoyer, C., 2006. Solar energy resource management for electricity generation from local level to global scale 55–68.
Pinker, R., Laszlo, I., 1992. J. Appl. Meteorol. 31, 194–211.
Polo, J., Wilbert, S., Ruiz-Arias, J.A., Meyer, R., Gueymard, C., Suri, M., Martín, L., Mieslinger, T., Blanc, P., Grant, I., et al., 2016. Sol. Energy 132, 25–37.
Ren21, R., 2017. Renewable Energy Policy Network for the 21st Century, Paris, France.
Riihelä, A., Carlund, T., Trentmann, J., Müller, R., Lindfors, A.V., 2015. Remote Sens. 7, 6663–6682.
Skoplaki, E., Palyvos, J.A., 2009. Sol. Energy 83, 614–624.
Stoffel, T., Renné, D., Myers, D., Wilcox, S., Sengupta, M., George, R., Turchi, C., 2010. Concentrating solar power: best practices handbook for the collection and use of solar resource data (CSP). Technical Report. National Renewable Energy Laboratory (NREL), Golden, CO.
Stroeve, J., Markus, T., Boisvert, L., Miller, J., Barrett, A., 2014. Geophys. Res. Lett. 41, 1216–1225.
Urraca, R., Gracia-Amillo, A.M., Koubli, E., Huld, T., Trentmann, J., Riihelä, A., Lindfors, A.V., Palmer, D., Gottschalg, R., Antonanzas-Torres, F., 2017. Remote Sens. Environ. 199, 171–186.
Wang, K., Augustine, J., Dickinson, R.E., 2012. J. Geophys. Res.: Atmos. 117.
Wiscombe, W.J., Warren, S.G., 1980. J. Atmos. Sci. 37, 2712–2733.

Paper III [redacted]

Accepted in Solar Energy with minor revision (under-review)

Solar radiation estimation at high latitudes: Assessment of the CMSAF databases, ASR and ERA5

Bilal Babar*, Rune Grand Graversen and Tobias Boström

Department of Physics and Technology, The Arctic University – University of Tromsø, Norway

*corresponding author: bilal.babar@uit.no

Abstract

There is a growing demand for the estimation of solar energy potential at high latitude locations. This study compares four datasets; Cloud, Albedo, Radiation dataset Edition 2 (CLARA), Surface Solar Radiation dataset – Heliosat Edition 2 (SARAH), ECMWF Reanalysis 5 (ERA5) and Arctic System Reanalysis v2 (ASR) on high latitude locations. Global horizontal irradiance (GHI) from these datasets is compared with *in-situ* ground-measurements over multiple locations in Norway. The first two datasets are mainly based on satellite estimation of solar radiation, while the latter two are based on a combination of a weather-prediction model, satellite data, and other observations. The datasets are evaluated against quality-controlled *in-situ* measurements of solar radiation from pyranometers. Overall, CLARA, SARAH, and ERA5 show moderate errors, while those of ASR are considerably larger. Monthly averages of global horizontal irradiance have mean absolute deviation (MAD) of 6.3 Wm^{-2}, 5.8 Wm^{-2}, 6.4 Wm^{-2}, and 14.5 Wm^{-2} for CLARA, SARAH, ERA5, and ASR, respectively. Seasonal error analysis of these datasets reveals that SARAH has the lowest errors in all seasons. The datasets are classified into clear-sky, intermediate-cloudiness, and overcast categories, by using two thresholds of cloudiness based on the ratio of radiation at ground to its corresponding clear-sky value (clear-sky index). The categories obtained from satellite and reanalysis data are then compared against estimates based on corresponding *in-situ* observations; this analysis shows that both CLARA and SARAH perform better than ERA5 and ASR for these categories. SARAH and CLARA perform similarly in all types of conditions, but a gradual increase in errors for an increase of cloudiness is observed for ERA5 and ASR. Yearly energy analysis shows that CLARA performs better than other datasets for locations above latitude 65ºN, and SARAH performs better in locations below 65ºN. A further analysis is performed to assess the cloud sensing abilities of ERA5. On a shorter time scale, there are errors due to inaccurate representation of clouds, however on longer time scales *i.e.* months and years, these errors are considerably reduced. ERA5 is observed to overestimate TCWC (the total cloud water content defined as the mass of water and ice in a cloud) in clear-sky and intermediate-cloudy categories, while in overcast category it is underestimated. Generally, an overestimation of solar radiation is observed in reanalysis and an underestimation is observed in satellite methods.

Keywords: Solar radiation, Arctic, Reanalysis, Satellite estimations, CMSAF, ECMWF

1. Introduction

Accurate solar resource measurements at potential photovoltaic (PV)/thermal installation sites are usually not available. For example, only a few meteorological stations record high-quality measurements in Norway (Øyvind et al., 2013). The assessment of solar resource at a specific location forms the basis for future installations. However, solar radiation is intermittent in nature and its variation on longer times scales is important for the planning of future installations (Crabtree et al., 2011). In addition, such information is also used in the long and short-term forecasting of power production and in optimizing energy dispatch strategies (Heinemann et al., 2006; Remund et al., 2008). Long time series of global horizontal irradiance (GHI) is used in the energy sector as well as in meteorology, agriculture, and climate studies.

The three main components to consider before installing a solar energy system are site selection, annual output and temporal performance/operating strategy. These components are directly related to the resource potential of the site, and can be evaluated by analysing long-term historical data series. Often a typical meteorological year (TMY), which is derived from the historical data *e.g.* within the past 30 years, is used to assess site locations for feasibility (Hall et al., 1978). Recent studies like those of Huld et al. (2018) and Stoffel et al. (2010) have shown that TMY is not a good indicator for predicting solar radiation for a given year, but rather it represents typical estimates of the average long-term conditions. Sufficiently long historical records from ground-measurements are seldom available for a given location for constructing a reliable TMY. Therefore satellite estimations and reanalyses provide an alternative to the ground-measurements for these estimations (Stoffel et al., 2010).

Estimating surface solar radiation from the visible spectrum of sensors installed on satellites is a well-developed procedure (Cano et al., 1986; Gautier et al., 1980; Rigollier et al., 2004; Tarpley, 1979). However, the accuracy of these methods is lower than ground measurements, but the advantages of the satellite methods include large spatial and temporal coverage (Noia et al., 1993). Surface solar radiation estimated from geostationary satellites provide up to sub-hourly values on a few km grid resolution, while polar orbiting satellites provide up to daily average solar radiation. All geostationary satellites have a limited spatial coverage because these are positioned over the equator at 0º. In the case of Meteosat First Generation (MFG) and Meteosat Second Generation (MSG) geostationary satellites, they have a coverage of ±65º in latitude and longitude. At latitudes higher than these, they encounter a flat angle of view that decreases the spatial resolution and increases errors. Alternatively, polar orbiting satellites can be used at high-latitude locations, as they provide almost global coverage. The main shortcoming of polar orbiting satellites is low sensing frequency, which varies from twice daily at the equator to 14 times a day near poles (Pinker and Laszlo, 1992; Platt, 1983). Satellite-based solar radiation-estimation methods have high accuracy, but some studies like that of Gueymard (2011) and Ineichen (2014) have shown that large errors may exist. For uncertainties and known issues within the satellite-based solar radiation estimation techniques see Suri and Cebecauer (2014).

In Earth System Models (ESM) or reanalysis, solar radiation is often referred to as down-welling surface shortwave flux. There are a number of studies where reanalyses have been used to estimate solar radiation and power (Boilley and Wald, 2015; Juruš et al., 2013; Wild et al., 2015). However, an increase in bias with increasing latitude was observed in one of the studies (Yi et al., 2011). The main advantages of reanalyses include multi-decadal time series, worldwide coverage, and free-of-cost availability. Recently, it has been found that reanalysis-based irradiance estimates can be a useful supplement when satellite irradiance is not available (Bojanowski et al., 2014; Urraca et al., 2018), although, many studies have reported overestimations in reanalysis (Boilley and Wald, 2015; Kennedy et al., 2011; Wild, 2008).

The aim of this paper is to analyse four different datasets regarding their accuracy and provide a comparative analysis for high-latitude conditions. Two of these are based on satellite methods, a polar orbital Cloud, Albedo, Radiation dataset Edition 2 (CLARA 2), and a geostationary Surface Solar Radiation dataset – Heliosat Edition 2 (SARAH 2). The other two are based on a combination of a weather-prediction model and various types of observations; a global reanalysis; ECMWF Reanalysis 5 (ERA5), and a dynamical downscaling of such a reanalysis (ERA-interim); Arctic System Reanalysis v2 (ASRv2). The analysis is performed for Norway, which represents a complex topography and a large variation in latitudes ranging from 59º to 70ºN. Previously, CLARA-A1 and CLARA-A2 datasets have been compared for multiple locations in Norway and Sweden (Babar et al., 2018). It was found that the new edition of CLARA has less number of missing data points. However, CLARA-A2's new data points, which previously were missing in CLARA-A1, have high errors. These points mostly lie in the high latitude locations where a snow cover is frequent. Because of the difficulties in differentiating snow covers from clouds, such errors exist. Here we extend this work and the novelty lies in evaluating the above-mentioned datasets for GHI for high-latitude locations and providing an analysis of these datasets in different conditions. The datasets are evaluated for daily means, monthly means, yearly means, seasonal analysis, energy analysis, and performance in different sky categories. Daily and monthly averages are evaluated by dividing the locations in four groups, including above 65ºN, below 65ºN, coastal and inland regions. In the final section, the effects of clouds in ERA5 are computed for different sky categories and compared with ground-measured solar radiation, which gives an insight into the challenges of estimating solar radiation in ERA5.

This paper is formatted as follows: Section 2 gives a description of the datasets analysed in this study. Section 3 provides an overview of the quality control procedures applied on the ground data and validation metrics. Section 4 presents the results and provides a brief discussion. Section 5 concludes the findings of this work.

2. Datasets

The datasets analysed in this study have different spatial and temporal resolution. Table 1 shows an overview of the datasets. SARAH and ASR can be considered as high-resolution datasets, while CLARA and ERA5 are coarse resolution datasets. SARAH is the highest resolution dataset with hourly temporal resolution and a spatial resolution of 0.05ºx0.05º. ASR contains data with three-hour temporal resolution and a spatial resolution of 15 km (0.136º). For both of these datasets, the nearest grid point from the site location is selected for data extraction. However, CLARA and ERA5 provide data on a much coarser grid of 0.25ºx0.25º and 0.28ºx0.28º, respectively. Data extraction from these datasets is performed by selecting the four surrounding grid points at site locations and applying inverse weighted-distance interpolation to obtain solar radiation at the coordinates of the site. In case of

CLARA, there are missing data points, which imply that at some of the periods there are no available data in the surrounding four points. When the surrounding points have less than three valid values, the interpolation is replaced by a missing value indicating that no valid values exist at that particular time and place. ASR and ERA5 do not contain missing values.

The datasets used in this study have certain spatial and temporal limitations. SARAH is limited to ±65º in latitude and longitude due to the shape of the viewing disc of MFG/MSG satellites and because of the flat viewing angle of geostationary satellites that results in increased errors above 65ºN. The evaluation of SARAH dataset is performed for locations below 65ºN latitude. CLARA and ERA5 are global datasets, whereas ASR is regional but covers all locations analysed in this study. SARAH and CLARA are available from 1983 to 2015 and 1982 to 2015, respectively. At the time of writing, ERA5 is available from 2000 to 2017. The years from 2000 to 2015 are included in this study from these datasets. ASRv2 is available from 2000 to 2012 and its complete available time series is used.

Table 1

Description of the datasets used in this study. The period analysed, spatial and temporal resolutions are shown for each dataset.

	Method	Years analysed	Spatial resolution	Highest temporal resolution	Spatial limits
CLARA -A2	Polar-orbiting Satellite	2000-2015	0.25ºx0.25º	24 Hours	Global
SARAH -2	Geostationary Satellite	2000-2015	0.05ºx0.05º	0.5 Hour	Limited to ±65º latitude and ±65º longitude
ERA5	Reanalysis (Global)	2000-2015	0.281ºx0.281º	1 Hour	Global
ASRv2	Reanalysis (Regional renalaysis downscaled from ERA-interim)	2000-2012	0.136ºx0.136º	3 Hours	180W - 180E longitude 24.643N - 90N latitude

2.1 CLARA-A2

The CLARA-A2 dataset was released in December 2016 and it is the second edition of CLARA (Cloud, Albedo, Radiation dataset) by satellite application facility on climate monitoring (CM-SAF). The dataset is available from 1 January 1982 to 31 December 2015, and constitutes an extension of 6 years relative to the previous CLARA-A1 dataset. This dataset has global coverage with a spatial resolution of 0.25ºx0.25º on a regular lat-lon grid and it provides daily and monthly averages of surface incoming shortwave radiation (SIS). To calculate daily averages, at least 20 observations of incoming solar radiation in each grid box are required; similarly, 20 valid daily averages are required to generate monthly averages (Trentmann and Kothe, 2016). Along with SIS, CLARA also provides longwave up and down-welling surface radiation.

The fundamental method used in calculating surface solar irradiance from satellite observations is based on the reflectance measured by the satellite instruments, which is related to the atmospheric transmittance. The underlying algorithm in CLARA uses the Advanced Very High Resolution Radiometer (AVHRR) sensor data to derive the atmospheric transmittance, which is used in calculating surface incoming solar radiation. The solar radiation is estimated by using the solar zenith angle, cloud coverage, vertically-integrated water vapour, and aerosol optical depth. Finding solar zenith angles is straightforward and can be calculated accurately. The vertically-integrated water vapour and aerosol optical depth are not available in the AVHRR data and for these fields, external sources are used. For vertically-integrated water vapour, ERA-Interim Reanalysis (Dee et al., 2011) is used and the vertical ozone column is set to a constant value of 335 DU, as its variability has negligible impact on the estimated solar radiation. Aerosol information is taken from the modified version of the monthly mean aerosol fields from Global Aerosol Data Set/ Optical Properties of Aerosols and Cloud (GADS/OPAC) climatology. In the algorithm, AVHRR data is used to retrieve only the cloud cover information. The first step in estimating surface solar radiation is the classification of the sky condition. The Nowcasting SAF (SAFNWC) software is used to derive the information on cloud coverage for each pixel by using the information from the satellite sensors. If no cloud is detected (cloud free pixel), surface solar radiation is calculated by using the clear-sky Mesoscale Atmospheric Global Irradiance Code (MAGIC) by using only auxiliary sources. If the pixel is classified as cloudy (cloud contaminated or fully cloudy), visible channels of AVHRR instrument are used to derive broadband reflectance. These reflectances are then transferred to broadband fluxes by using a bidirectional reflectance distribution function (BRDF). In the next step, these broadband top-of-the-atmosphere albedos are used to derive transmissivity through a look-up table approach. Finally, the transmissivity is used in calculating surface solar radiation. In this

dataset, all data points with a solar zenith angle larger than 80º are set to missing values and solar zenith angle larger than 90º is set to zero. However, because a temporally constant surface albedo is used in the algorithm, this dataset does not provide radiation estimates on snow and sea ice coverage areas because changes in the albedo of the snow-covered surfaces are not considered (Karlsson et al., 2017). High-latitude locations may have a very different surface albedo than the temporally constant albedos considered in the algorithm. Such grid points are identified by calculating the difference between monthly mean CLARA-A2 SAL (surface albedo) data record and the surface albedo used in the processing of SIS. These critical grid points, which have a difference in surface albedo exceeding 35%, are masked-out from the final product by setting them as missing values. For more information on the CLARA dataset and its accuracy refer to Karlsson et al. (2017).

2.2 **SARAH-2**

The second version of surface solar radiation dataset – Heliosat (SARAH-2) is a climate data record of surface solar radiation by CMSAF (Pfeifroth et al., 2017a) and covers a period of 32 years from 1983 to 2015 and the region from 65ºN to 65ºS latitude and 65ºW to 65ºE longitude. The spatial resolution of the data is 0.05ºx0.05º (approximately 5km) and the data is available for 30 minutes instantaneous, hourly, daily, and monthly averages of surface incoming shortwave radiation on a horizontal surface, direct normal irradiance (DNI) and effective cloud albedo (CAL). To calculate daily averages at least three samples per day are required; similarly, 10 existing daily averages are required to generate monthly averages.

In this dataset, the broadband visible channels from Meteosat Visible Infra-Red Imager (MVIRI) instrument on-board the Meteosat first generation satellites and the Spinning Enhanced Visible and Infrared Imager (SEVIRI) instruments on-board the Meteosat second generation satellites are used to calculate the shortwave surface radiation. In the first step, the effective cloud albedo (CAL) is retrieved from the satellite data by using a modified Heliosat method (Hammer et al., 2003). This modification of the Heliosat method in combination with gnu-MAGIC/SPECMAGIC is called MAGICSOL. The modified Heliosat method provides the broadband effective CAL, but to consider the spectral effect of clouds, a Radiative Transfer Model (libRadtran) based correction is applied. The CAL is related to the cloud transmission and, hence, by calculating clear-sky radiation, the all-sky radiation can be estimated. In this dataset, for calculating clear-sky radiation the SPECMAGIC model is used, which is based on a so-called hybrid eigenvector look-up table approach (Mueller et al., 2012). The input parameters for gnu-MAGIC/SPECMAGIC are date, time, solar zenith angle, coordinates, effective cloud albedo (cloud index), water vapour column density, surface albedo, aerosol optical thickness, and single scatter albedo for aerosols. Monthly mean values of vertically-integrated water vapour are taken from ERA-interim global reanalysis record (Dee et al., 2011), and monthly mean aerosol information is taken from Monitoring Atmospheric Composition and Climate project (MACC) aerosol climatology. Surface solar radiation is derived from combining the SPECMAGIC algorithm and the effective cloud albedo (Pfeifroth et al., 2017b). One of the limitation of SARAH is that for solar zenith angles between 88º and 90º, the corresponding data points are set as missing values, and above solar zenith angle of 90º, the data points are set to zero. Improvements in the new version of the dataset includes stability during the change of instrument from MVIRI to SEVIRI in 2006, and correction of the cloud albedo to account for the slant viewing geometry effects (Pfeifroth et al., 2018). For more information on the retrieval methods refer to Müller et al. (2015).

2.3 ERA5

ECMWF Reanalysis 5 (ERA5), is the fifth generation of European Centre for Medium-Range Weather Forecasts (ECMWF) atmospheric reanalysis of the global climate and span a period of 1950 to near real time (Hans and Dick, 2016). At the time of this study, data from 2000 to 2017 are available. Further data back in time will be released in 2019-20 and will continue to update forward in real-time. In ERA5, the solar radiation variable has a spatial resolution of 31km (0.28125ºx0.28125º) and an hourly temporal frequency. ERA5 uses the Integrated Forecasting System (IFS) cycle 41r2 with a state-of-the-art four-dimensional variational analysis (4DVAR) assimilation system. ERA5 has more pressure levels than ERA-Interim (the previous edition of ECMWF reanalysis) and more variables are made available for this reanalysis than for those of earlier generation. For more information on ERA5 refer to ECMWF (2018).

In this study, shortwave surface downward radiation, shortwave surface downward radiation clear-sky, and total cloud water content (the vertically-integrated cloud water concentration) are used from this dataset. In ERA5, the incoming short wave radiation is obtained from a Radiative Transfer Model (RTM). This model simulates the attenuation in solar radiation caused by the atmosphere, therefore, the quality of estimated radiation depends on

the RTM used. Reanalysis generally do not assimilate aerosol, clouds or water vapour data, which increases the uncertainty in the estimated surface irradiance (You et al., 2013; Zhao et al., 2013).

2.4 Arctic System Reanalysis v2

Arctic system reanalysis version 2 (ASRv2) is a polar-optimized dynamic downscaling of ERA-Interim reanalysis by using Weather Research and Forecast Model (WRF) version 3.6.0. The data set is available for the period of 2000 to 2012. The grid resolution is 15km, which is finer than most global models and the previous release of ASR (ASRv01), whereas the time resolution of the dataset is 3 hours. The downscaling is optimized for Polar Regions, and polar physics is used where possible, including heat transfer through snow and ice, the fractional sea ice cover, the ability to specify variable sea ice thickness, snow depth on sea ice and sea ice albedo, as well as other optimizations included in the Noah Land Surface Model. The area covered by this dataset is 1.2×10^8 km^2, which is about 50% of Northern hemisphere. Spectral nudging from ERA-Interim is applied on geopotential height, temperature, and wind components above 100 hPa on the inner domain. ASR uses three-dimensional variational analysis (3DVAR) for observations, including radiance data, from a number of satellites (Bromwich et al., 2017).

3. Ground data

In this study, 31 locations from Norway are analysed for the four mentioned datasets. The coordinates of the locations, altitudes, and land type are indicated in appendix A and an overview of site locations is shown in Figure 1. The ground-measured data is acquired from the Norwegian Institute of Bioeconomy Research (NIBIO). NIBIO registers hourly-average GHI by using Kipp and Zonen CMP11 or CMP13 pyranometers. The data is quality controlled and the equipment is maintained regularly on a daily or weekly basis (http://lmt.bioforsk.no/about). The daily averages of ground data were calculated by following Urraca et al. (2017b), where these were calculated for those days when at least 20 valid hourly means were available, however when this criteria was not met the daily average was replaced by a missing value. Similarly, the monthly averages were calculated for those months when all the hourly values were available. If this condition was not met, the monthly average was replaced by a missing value (Roesch et al., 2011). The amount of missing data in the ground measurement was largely reduced because of the application of quality control procedures (explained in the next section).

In this study, the numbers of years used from each data set are different. For ASR, 12 years of data is used and 16 years of data is used for ERA5, CLARA, and SARAH. Furthermore, the sites are divided four groups; above 65ºN, below 65ºN, inland and coastal regions. The studied locations are divided into coastal and inland regions are grouped by observing the proximity to the shoreline. Regions within 30 km of the shoreline are considered as coastal. From the 31 locations studied here, 14 sites are classified as coastal and 17 sites as inland, while 4 sites lie above 65ºN and 27 lie below 65ºN latitude. For details on the land-type classification, refer to appendix A.

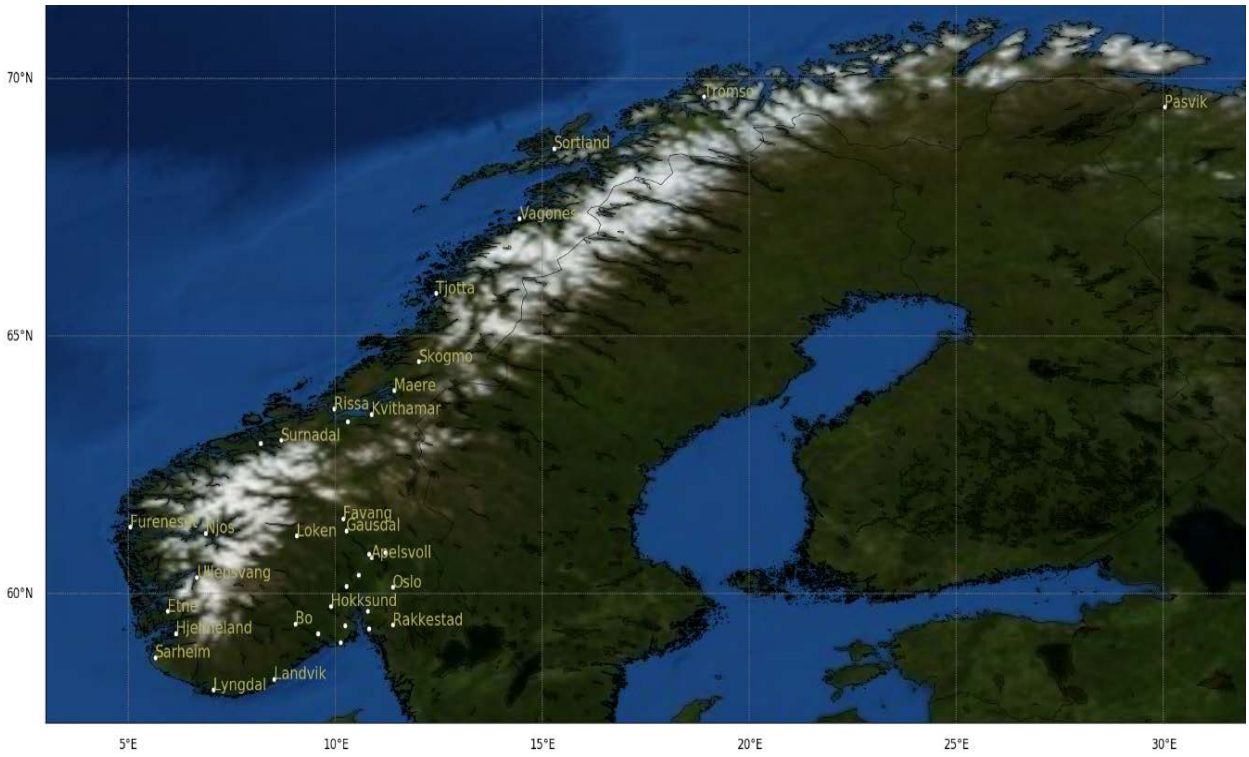


Figure 1: Locations of the sites included in the study. To avoid overlapping of names some locations are shown with only white dots.

3.1 Quality Control

Although the data provided by NIBIO is quality controlled, Urraca et al. (2017a) observed that operational and equipment errors exist in NIBIO stations. The first quality-control check performed in this study is to look at the percentage of missing data. Any year having more than 5% of missing values is discarded from the analysis. The second check is performed by using BSRN Global Network recommended Quality Control tests, V2.0 (Long and Dutton, 2010). These quality checks test values that are extremely rare and physically impossible. From this test, years having more than 1% of flagged values are removed from the ground data. The third quality control procedure is applied by using the Urraca et al. (2017a) quality control technique. In this test, CLARA and ERA5 datasets are used to check the quality of ground measurements by constructing confidence intervals to detect the operational and equipment errors. Following Urraca et al. (2017a), the locations in Norway are divided into two sections by grouping locations above 65ºN and locations below 65ºN. Separate confidence intervals are constructed for these groups of locations. After constructing these confidence intervals, the ground data is passed through an algorithm to check the data with errors, which appear in the form of flags. Following Urraca et al. (2017a) two checks are performed, one to see the operational errors and the other to see the equipment errors. After these checks, the years having large number of flags are visually inspected and removed from the analysis. Initially Pasvik, Mære, Njøs, and Ullensvang were included in the study but due to a large number of flags from the third quality control test, these were discarded. Pasvik and Ullensvang were found to have equipment errors and frosting, while Mære and Njøs were found to have shading errors. For more information on this quality control procedure refer to Urraca et al. (2017a). 2006 and 2007 were found to have a large number of missing data points; these were discarded from all locations. Gap filling methods are only used in calculating yearly energy averages by using nearest-neighbour interpolation. See appendix B for details about the years not included in the study. After performing quality control on the ground data, errors might still exist but in addition to validating the datasets, this study provides a comparative analysis of these datasets for high latitude locations. From a comparative point of view, the errors in the ground data will have a similar effect on all datasets.

3.2 Validation

In order to evaluate the performance of the datasets, some common statistical measures are used. The most widely used measure is the root mean square deviation (RMSD). As an addition, the BIAS or mean bias deviation (MBD) is used in the evaluation. MBD gives an insight in under or over estimations. Mean absolute deviation (MAD) is

also used for the evaluations. Because of the absolute values used in this measure, the negative and positive errors do not cancel out as in the BIAS. MAD is a good measure for comparing different models. Moreover, Pearson correlation and scatter plots are used to indicate the spread and overall correlation of the datasets with ground measurements.

4. Results and discussion

Table 2 lists the RMSD, MAD, and MBD of the datasets for the locations included in the study. The error indicators in table 2 are expressed in Wm^{-2} and values in parentheses are daily averages. Night-time values are included in calculating daily and monthly averages. Along with all sites included in the study, table 2 also shows error metrics for above 65ºN, below 65ºN, inland and coastal regions.

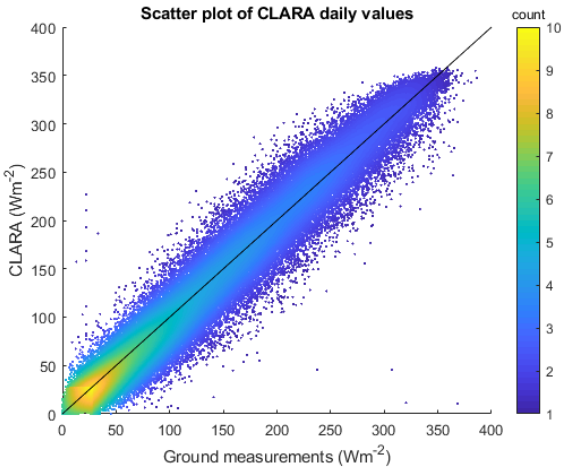
Table 2
Error metrics expressed in Wm^{-2}, for the datasets analysed in this study. Numbers without parentheses are monthly averaged errors while those in parentheses are daily averaged errors. Numbers are averaged over all stations. Error metrics for different geographical groups are also shown.

	RMSD(Wm^{-2})				MAD(Wm^{-2})				MBD(Wm^{-2})			
	CLARA	SARAH	ERA5	ASR	CLARA	SARAH	ERA5	ASR	CLARA	SARAH	ERA5	ASR
All Sites	9.5 (18.3)	8.7 (18.0)	9.9 (26.4)	21.7 (42.6)	6.3 (12.8)	5.8 (11.8)	6.4 (16.7)	14.5 (27.1)	-3.0 (-1.7)	-3.6 (-2.5)	2.1 (4.0)	13.1 (16.9)
Above 65ºN	10.1 (16.0)	-	10.9 (26.3)	20.3 (39.4)	5.3 (9.7)	-	6.1 (14.5)	11.1 (21.5)	-3.4 (-2.8)	-	3.8 (5.6)	8.0 (11.0)
Below 65ºN	9.4 (18.6)	8.7 (18.0)	9.9 (26.8)	21.9 (43.0)	6.5 (13.2)	5.8 (11.8)	6.5 (17.3)	15.0 (27.9)	-3.0 (-1.5)	-3.6 (-2.5)	2.0 (4.0)	13.8 (17.8)
Coastal	9.1 (17.5)	8.5 (17.1)	10.0 (26.5)	21.8 (41.9)	5.9 (12.1)	5.6 (11.2)	6.2 (16.3)	13.9 (25.6)	-2.7 (-3.1)	-3.4 (-2.2)	2.3 (4.3)	11.9 (15.7)
Inland	9.3 (23.4)	8.8 (18.4)	10.0 (26.9)	21.7 (43.1)	6.2 (14.5)	5.9 (12.1)	6.7 (17.6)	15.0 (28.3)	-3.0 (-5.0)	-3.7 (-2.7)	2.2 (4.1)	14.0 (18.0)

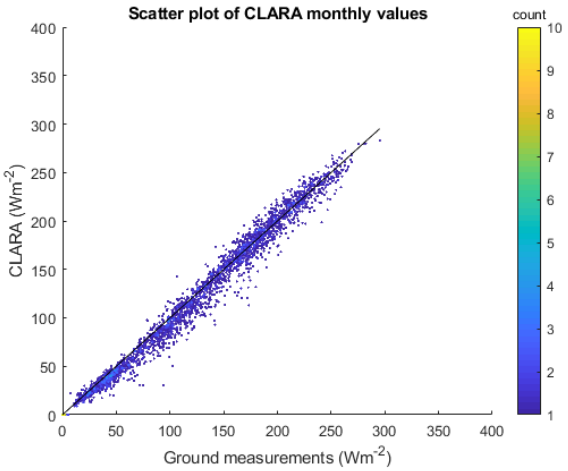
From the table it can be seen that for all locations, SARAH provides the best estimation in terms of RMSD, while ASR performs the worst. The same pattern follows on the MAD errors where SARAH performs better than other datasets, while ASR has the highest errors. ERA5 and ASR (reanalysis models) are observed to be overestimating, similar to previous studies (Boilley and Wald, 2015; Kennedy et al., 2011; Wild, 2008). Both CLARA and SARAH (satellite databases) underestimate solar radiation (Posselt et al., 2012; Riihelä et al., 2015). At slant angles of view, such as those experienced by geostationary satellites at high latitudes, solar radiation is often underestimated by satellite methods because of an overestimation in cloud. The highest bias is seen in ASR while biases of CLARA, SARAH, and ERA5 are very similar in magnitude.

The table also shows RMSD, MAD, and MBD for location categories above 65ºN, below 65ºN, coastal and inland. Above 65ºN latitude, CLARA has the lowest errors and ASR has the highest errors while ERA5 provides moderate errors. SARAH does not provide coverage above 65ºN latitude. At locations below 65ºN, SARAH and CLARA have low errors as compared to other datasets. The ASR has the highest errors at such locations as well. SARAH and CLARA have lower errors in coastal regions than inland, mainly due to less snow covers in coastal regions (Babar et al., 2018). Note that CLARA and ERA provide data at a similar spatial resolution, i.e. 0.25º and 0.28º, however the surface radiation in CLARA is calculated at much finer resolution (around 4km) than in ERA5, therefore, CLARA performs better at coastal regions. On the contrary, in inland locations SARAH provides better estimates than other datasets. CLARA comes second in terms of both daily and monthly means, while ASR performs the worst. In this analysis, ERA5 is seen to perform better at locations below 65ºN than above 65ºN latitude.

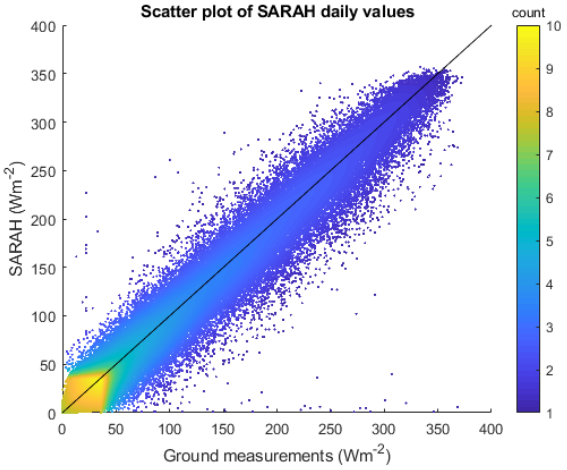
Figure 2 (a-h) illustrates the scatter plots of the monthly and daily averages of the datasets. The black coloured line represents the x=y line for reference. Evidently, CLARA and SARAH have a very similar spread on both monthly and daily averages. A correlation of 0.98 for daily means and 0.99 for monthly means are observed for both of these datasets. ASR has a wider spread in scatter plots with correlation coefficients of 0.99 and 0.92 for monthly and daily means respectively. In addition, a positive bias in ASR monthly averages can be observed. ERA5 has an intermediate spread with a correlation of 0.99 for monthly averages and 0.95 for daily averages.


Scatter plot of CLARA daily values
CLARA (Wm-2)
Ground measurements (Wm-2)
count

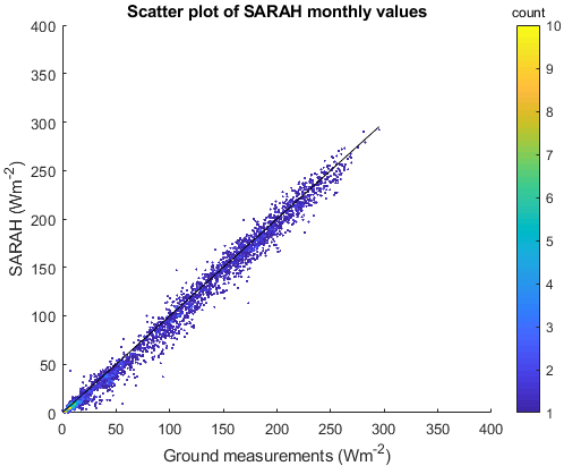
(a)


Scatter plot of CLARA monthly values
CLARA (Wm-2)
Ground measurements (Wm-2)
count

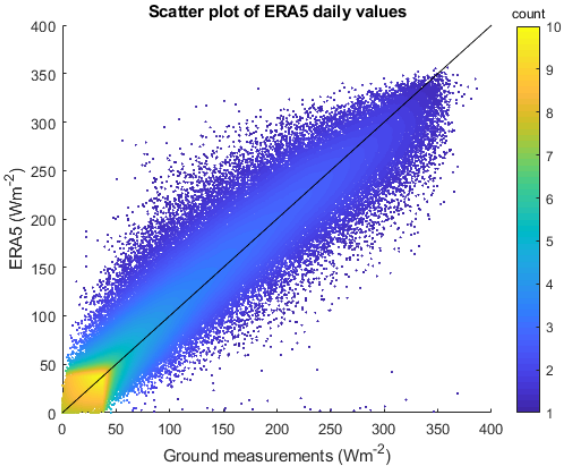
(b)


Scatter plot of SARAH daily values
SARAH (Wm-2)
Ground measurements (Wm-2)
count

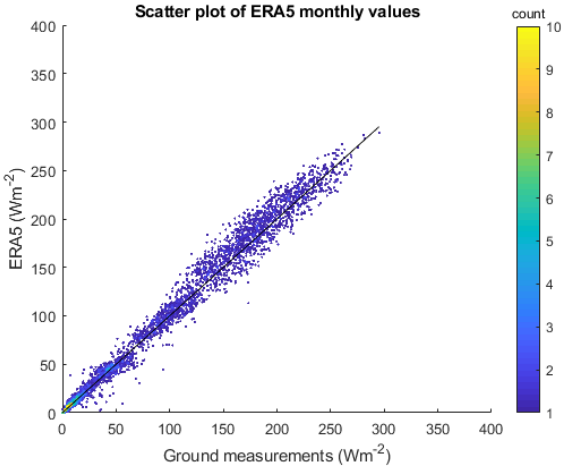
(c)


Scatter plot of SARAH monthly values
SARAH (Wm-2)
Ground measurements (Wm-2)
count

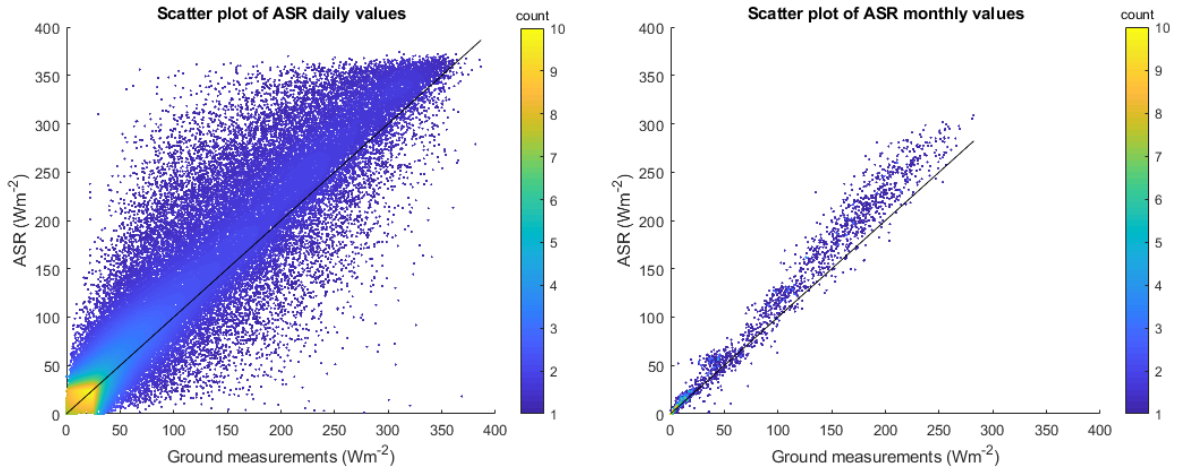
(d)


Scatter plot of ERA5 daily values
ERA5 (Wm-2)
Ground measurements (Wm-2)
count

(e)


Scatter plot of ERA5 monthly values
ERA5 (Wm-2)
Ground measurements (Wm-2)
count

(f)



(g) (h)

Figure 2: Monthly mean and daily mean GHI scatter plots of the datasets. Ground *in-situ* observations and estimated values of solar radiation are given in Wm^{-2}. The legend bar shows the density of data points on a coloured scale. Satellite data show narrow spread and underestimation while reanalyses show wider spread and overestimation.

Table 3

Statistical errors of the yearly average energy estimates for the datasets in kWh per square meter and year on a horizontal surface. Energy statistics for different geographical groups are also shown.

	Energy ($KWh.m^{-2}.year^{-1}$/percentage error)											
	CLARA			SARAH			ERA5			ASR		
	Est.	Obs.	%Err.	Est.	Obs.	%Err.	Est.	Obs.	%Err.	Est.	Obs.	%Err.
All Sites	838.4	862.9	-2.8	861.2	880.5	-2.2	908.1	862.9	+5.2	1017.1	865.5	+17.5
Above 65ºN	711.7	715.5	-0.5	-	-	-	806.0	715.5	+12.6	870.4	751.3	+15.9
Below 65ºN	853.5	880.5	-3.1	861.2	881.2	-2.3	920.3	881.2	+4.4	1034.8	879.3	+17.7
Coastal	845.4	857.6	-1.4	882.2	899.9	-1.9	904.9	857.6	+5.5	1009.4	862.6	+17.0
Inland	832.3	867.5	-4.1	847.1	867.5	-2.4	911.0	867.5	+5.0	1023.2	867.9	+17.9

In addition to daily and monthly errors, energy stakeholders use the yearly solar radiation energy averages to evaluate the existing energy systems and plan new projects. Estimated yearly radiation gives an insight into the total production of such systems and can be compared with the yearly consumption to increase efficiency of such systems. Table 3 shows yearly average energy outputs in terms of estimated, observed and percentage error. The yearly energy averages were calculated by integrating the daily averages of the datasets. The gaps in CLARA, SARAH and ground-measured data are filled by using nearest-neighbour interpolation. The SARAH performs better than other datasets, but with CLARA following just behind. Above 65ºN, CLARA gives much lower deviations than ERA5 and ASR, while SARAH has no coverage. It can be observed from the table that ERA5 performs better at inland locations while other datasets perform better at coastal regions. It has been documented that satellite estimation methods deteriorate over snow-covered surfaces. In Norway, usually inland locations have a higher snow-depth than the coastal regions. Because of the shortcoming of satellite estimation algorithm in the differentiation of clouds from snow covers, satellite-based data do not perform as well in snow-covered areas as on snow-free areas. However, both satellite-based datasets underestimated the energy as shown by a previous study (Babar et al., 2018), while the reanalyses are observed to be overestimating. ERA5 overestimates the energy production much more at locations above 65ºN than below; other datasets give very similar deviations in energy averages at different locations. The results of this analysis shows that below 65ºN latitude, the SARAH performance is better than that of the other data sets. In addition to higher spatial and temporal resolution, the errors in this dataset are low. Above 65ºN, only CLARA gives reasonable errors.

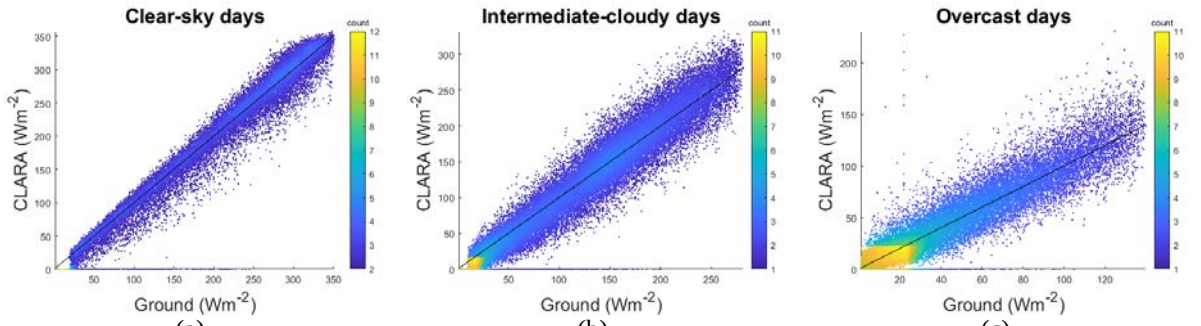
Analysis of yearly averaged GHI in terms of RMSD, MAD and MBD is shown in appendix D. For the yearly averages, high errors are observed in ASR when all locations are taken into account, while CLARA, SARAH, and ERA5 give considerably lower errors. CLARA is observed to perform better at coastal locations than in the inland regions, while the errors increase at locations above 65ºN. SARAH has no coverage above 65ºN, and the deviations are larger at inland regions than at the coast. ERA5 provides similar errors as those of CLARA in inland, above

65ºN and below 65ºN, but shows high errors in coastal regions. CLARA performs better than ERA5 at coastal regions, because the surface radiation calculation in CLARA is made at a much finer resolution (0.05º) than in ERA5, and therefore, takes into account the changing surface conditions of the coastal regions to a larger degree. ASR on the other hand gives the highest errors among the datasets for all locations.

A seasonal analysis of the datasets is performed by dividing a typical year into 4 parts, where February to April are grouped in FMA, May to July are grouped in MJJ, August to October are grouped in ASO and November to January are grouped in NDJ. This division into seasons is made so that summer solstice is approximately in the middle of the summer season. Table E1 in appendix E illustrates the seasonal error analysis of the datasets and it shows that the RMSD values are high in FMA, and decreases as the year progresses. ASR is observed to have high monthly and daily RMSD. MAD values in the table show that monthly mean values are similar for CLARA, SARAH, and ERA5 while ASR gives considerably larger MAD. MBD shows that both reanalyses overestimate solar radiation and satellite methods mostly underestimate it. In this analysis, SARAH, CLARA, and ERA5 perform similarly and better than ASR. Moreover, there are larger errors in satellite methods than reanalyses in FMA and MJJ, mostly because of the presence of snow covers, which are difficult to differentiate from clouds in such methods (Babar et al., 2018). Low solar elevation angles at high latitude locations make this differentiation further challenging. On the contrary, ERA5 performs better than satellite datasets in FMA and NDJ at high latitudes. However, the performance of satellite methods improves in summer and autumn months.

4.1 Evaluation of different sky conditions

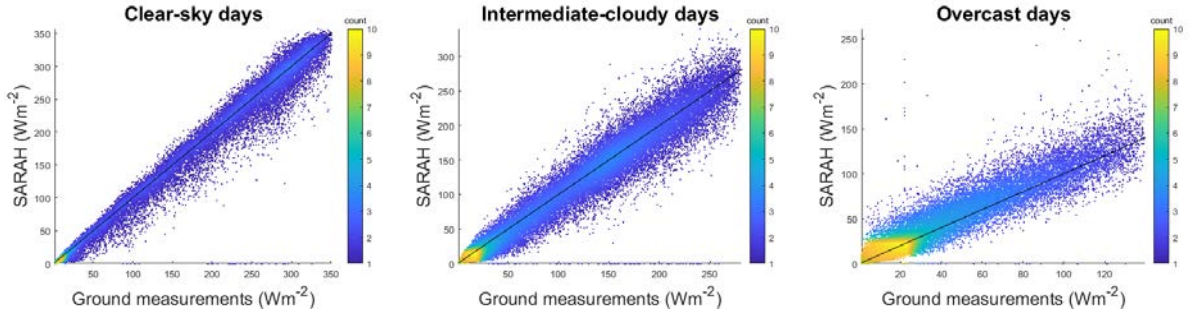
To evaluate the datasets for their performances in different sky conditions, the datasets were divided into clear-sky, intermediate-cloudiness, and overcast categories. This division is established based on the clear-sky index (*Kc*), which is defined as the ratio of GHI recorded on the ground to the clear-sky GHI. The BIRD clear-sky model is used to calculate the clear-sky values at the ground measurement locations (Bird and Hulstrom, 1981). After calculating clear-sky index, *Kc*, following Smith et al. (2017) and Widén et al. (2017), values higher than 0.8 are considered indicating a clear-sky day, values of *Kc* between 0.4 and 0.8 are considered as intermediate-cloudy and values below 0.4 are considered as overcast.



(a) (b) (c)

CLARA	RMSD (Wm^{-2})	MAD (Wm^{-2})	MBD (Wm^{-2})
Clear-sky	21.5	13.8	-4.0
Intermediate-cloudiness	22.1	16.0	-3.3
Overcast	12.8	8.7	-0.2

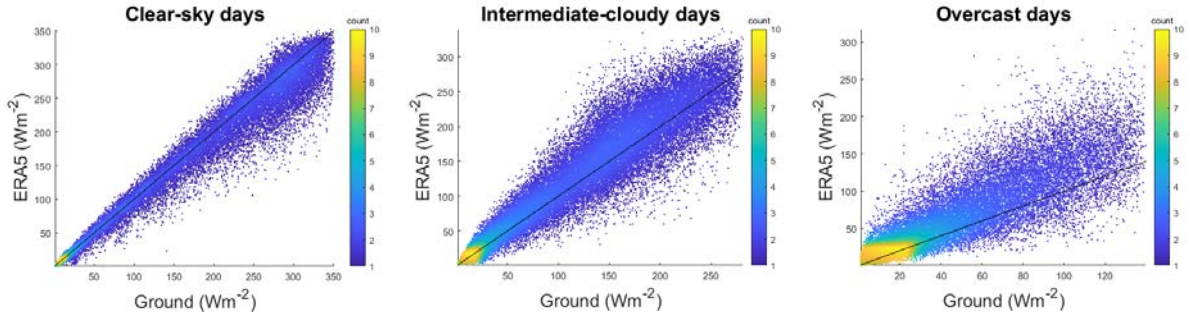
Figure 3: CLARA daily averaged errors under clear-sky, intermediate-cloudiness, and overcast categories. Scatter plots for the different sky-categories are shown. The coloured legend bar shows the density of points in the scatter plot.



(a) (b) (c)

SARAH	RMSD (Wm^{-2})	MAD (Wm^{-2})	MBD (Wm^{-2})
Clear-sky	20.4	12.8	-5.5
Intermediate-cloudiness	20.2	13.5	-3.0
Overcast	13.2	8.7	4.4

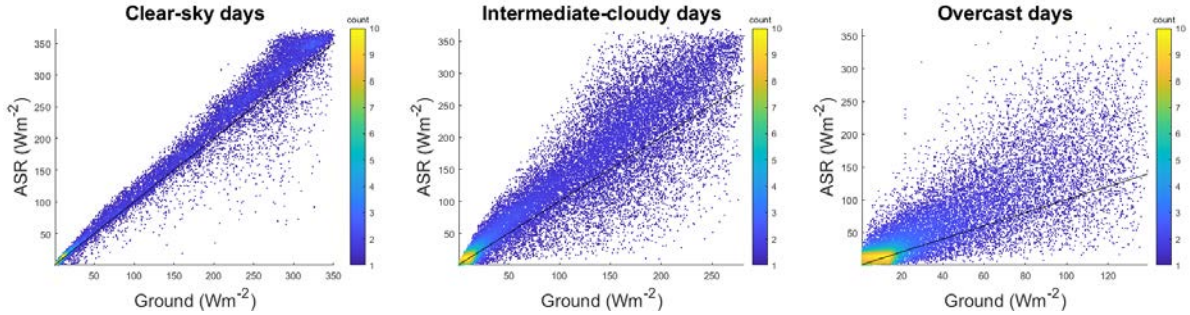
Figure 4: As Figure 3, but for SARAH.



(a) (b) (c)

ERA5	RMSD (Wm^{-2})	MAD (Wm^{-2})	MBD (Wm^{-2})
Clear-sky	25.5	16.8	-10.0
Intermediate-cloudiness	28.5	19.9	8.7
Overcast	29.6	18.6	15.2

Figure 5: As Figure 3, but for ERA5.



(a) (b) (c)

ASR	RMSD (Wm^{-2})	MAD (Wm^{-2})	MBD (Wm^{-2})
Clear-sky	29.2	21.1	11.6
Intermediate-cloudiness	51.3	37.2	23.3
Overcast	49.0	30.8	25.0

Figure 6: As Figure 3, but for ASR.

Figure 3-6 show the results of cloudiness classification of the datasets. Overall in the three categories, SARAH performs better than other datasets while ASR performs the worst. In clear-sky category, an underestimation is observed in SARAH, CLARA, and ERA5, while ASR overestimates radiation. CLARA performs slightly worse than SARAH in this category, but both have the same correlation coefficients of 0.98, while ERA5 and ASR both have a correlation of 0.97. Similarly, in the intermediate-cloudy category, both satellite databases underestimate, while reanalyses overestimate. Finally, in the overcast category, CLARA slightly underestimates solar radiation while other datasets overestimate. In this category, SARAH and CLARA are found to perform very similar with correlation coefficients of 0.95 and 0.94, respectively. It should be noted that the sky cloudiness differentiation is

performed on the basis of a clear-sky model and ground observed GHI. In conclusion, all the models have discrepancies in presenting clouds in all types of sky conditions.

As explained in Section 2, under clear-sky conditions CLARA uses aerosol information from Global Aerosol Data Set/Optical Properties of Aerosols and Clouds (GADS/OPAC) climatology and integrated water-vapour information from ERA-interim, and SARAH uses both Monitoring Atmospheric Composition and Climate (MACC climatology) and integrated water-vapour from ERA-Interim. Aerosol information from MACC climatology is observed to have higher accuracy than GADS/OPAC climatology (Mueller and Träger-Chatterjee, 2014). The maximum aerosol optical depth (AOD) is reduced in GADS/OPAC climatology for the CLARA dataset, but the results show that the climatology used in SARAH performs better than in CLARA even after the modifications. The negative biases observed in the clear-sky and intermediate-cloudy categories are possibly due to incorrect prediction of clouds and the aerosol climatology being too thick, which results in an underestimation of solar radiation. As reported in Mueller and Träger-Chatterjee (2014) and Polo et al. (2014), both MACC and GADS/OPAC climatologies result in underestimation of surface solar radiation because of the apparent overestimation in AOD thickness. In addition to aerosol optical depth, vertically-integrated water vapour values taken from ERA-Interim are shown to be too large (Kishore et al., 2011), which can further attenuate the surface solar radiation. Moreover, monthly mean values of aerosol optical depths are used which might also cause errors for daily resolutions. In ERA5, the radiative transfer model RTTOV11 (Radiative Transfer for TOVS) has a tendency to underestimate reflectance of high cumulus cloud tops while the reflectance of lower water clouds is overestimated. These cloud top reflectance errors possibly result in an underestimation in clear-sky conditions and overestimation in intermediate-cloudy and overcast conditions. In ASR, all the conditions are overestimated which shows that there is an underestimation in aerosol optical depth and cloudiness in the atmosphere.

After analysing different sky conditions, it can be concluded that estimations based solely on satellite retrievals generally provide a much better result. However, SARAH is limited to 60-65ºN (in Scandinavia) and CLARA is limited to daily and monthly means. For high latitude and high recording frequency, ERA5 can still provide an alternative, especially for clear-sky and intermediate-cloudy conditions in cases where satellite coverage is not available or have missing data.

4.2 Analysis of daily average TCWC and daily sky-condition classification in ERA5

To analyse the cloud placement of ERA5, the total cloud water content (TCWC) and short wave solar radiation downward, clear-sky (SWSDC) from ERA5 are used here. To obtain TCWC, total column liquid condensate and total column ice condensate from ERA5 were added together. ERA5 and other reanalyses have an overestimation or a positive bias in solar radiation as documented here and in accordance with Urraca et al. (2017b) and Urraca et al. (2018). On the contrary, satellite methods have a negative bias but higher accuracy (Riihelä et al., 2015). Reanalyses are based on weather-prediction models, and although assimilation of observations to some extent constrains these models, the weather patterns of the reanalysis may still be out of phase with reality. A small misrepresentation of clouds in space and time may have a large impact on the high-frequency correlation between model and *in-situ* observations, with regard to radiative fluxes such as solar radiation, and hereby large RMSD are induced. However at longer time scales, *i.e.* monthly or yearly time scales, the reanalysis may represent cloud frequency to a satisfactory degree because large errors in daily averages are compensated for in the seasonal mean, implying that reanalysis becomes a valuable alternative for estimating local solar resources. This can be observed by comparing the daily and monthly RMSD of ERA5 with satellite based datasets in table 2. For all the locations, the RMSD of monthly values for ERA5 is similar to that of CLARA and SARAH, but the RMSD of daily values (in parentheses) is considerably larger in ERA5 as compared to the satellite databases. On even longer time scales the difference decreases further, which can be observed by analysing yearly averages from table D1 in appendix D. In this section, the cloud representation in ERA5 on daily averages is explored (for years 2000 to 2015) and an analysis is given on the random errors in the presence of clouds at lower time scales. Clear-sky indices for all datasets are obtained by using SWSDC from ERA5 because the clear-sky values from ERA5 have the aerosol and water content information, which is used in calculating the surface solar radiation. The approach used in Section 4.1 is used here to classify days into the three categories by using clear-sky index, *Kc*. The analysis in this section is performed for days when the solar zenith angle is lower than 90º.

Table 4

The number of days and mean TCWC from *in-situ* ground measurements, ERA5 and CLARA are shown in the table for different sky categories. The number of days and mean TCWC in each cloudiness category for ERA5 is shown separately for cases when ERA5 and ground measurements agree on classification and for cases when there is a disagreement. Years from 2000 to 2015 are used in the analysis over all locations included in the study (see appendix B).

	Ground data		CLARA data		ERA5 data		ERA and ground agree		ERA and ground disagree	
	No. of days	Mean TCWC ($Kg.m^{-2}$)	No. of days	Mean TCWC ($Kg.m^{-2}$)	No. of days	Mean TCWC ($Kg.m^{-2}$)	No. of days	Mean TCWC ($Kg.m^{-2}$)	No. of days	Mean TCWC ($Kg.m^{-2}$)
Clear-sky	38265 (30.2 %)	0.03	39516 (31.3%)	0.03	53211 (33.4%)	0.02	29500	0.02	8765	0.07
Intermediate -cloudiness	49207 (38.8 %)	0.09	45244 (35.8%)	0.10	75268 (47.4%)	0.10	34700	0.10	14507	0.07
Overcast	39181 (30.9 %)	0.22	41417 (32.8%)	0.22	30389 (19.1%)	0.29	20914	0.30	18004	0.12

Table 4 shows the number of days and mean TCWC for each of the sky categories. In table 4, daily averages of solar radiation from CLARA are used to make a comparison with ERA5 in sky classification. It can be seen that ground measurement and CLARA classify almost the same percentage of days into each category even though the number of days available for these are not the same because of the missing values. CLARA also gives very similar mean TCWC values as ground measurements. On the contrary, ERA5 is observed to classify a higher number of days as intermediate-cloudy and a lower number of days as overcast than *in-situ* observations, hence showing that it has a negative bias towards classifying a day as overcast. Moreover, in ERA5 the mean TCWC is slightly underestimated in the clear-sky category but largely overestimated in overcast category. Table 4 further shows the number of days and mean TCWC for conditions when ERA5 and ground measurements agree on classification and for when there is a disagreement. Here it can be seen that the mean TCWC of days with agreement is the same as that of ERA5, but on the days of disagreement, there is an overestimation in mean TCWC in clear-sky days and an underestimation in overcast days. These results show that on clear-sky days, ERA5 has more clouds than *in-situ* observations, which is seen by higher levels of TCWC, while on the overcast days there are a lower amount of clouds, which is seen by lower levels of TCWC. However, it can be seen from the table that in clear-sky category, ERA5 and ground-measurements agree 77% of the time. The agreement on sky-condition is smaller in intermediate-cloudy category where 41% of the time ERA5 predicts the same conditions as *in-situ* observations, while the agreement in overcast category is 53%. Overall, 67.3% of the times it is seen that ERA5 and ground measurements classify the same conditions. Figure 7 shows the scatter plot of ground measurements and ERA5 for both of these conditions, when there is an agreement on classification and when there is a disagreement. It can be seen that the spread is large when there is a disagreement. A correlation coefficient of 0.98 is found for agreement data points while a correlation coefficient of 0.90 is found for disagreement point.

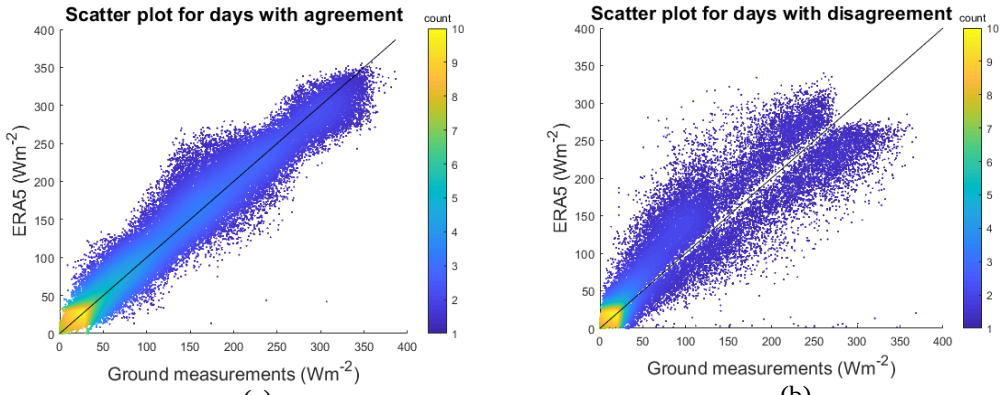


Figure 7: Scatter plots for the days when ERA5 and ground measurement agree in classification and when there is a disagreement. A correlation coefficient of 0.98 is found for agreement points and 0.90 for disagreement points.

Table 5 illustrates RMSD, MAD, and MBD of ERA5 in different sky categories. It shows the error metrics for the days when ERA5 and ground measurements agree on a category and for when there is a disagreement. The days of agreement on sky categories in table 5 can be compared to the deviations presented in Section 4.1, Figure 3. It can be seen that on the days of agreement ERA5 performs very similar to CLARA. However, large errors are observed when ERA5 does not agree with ground measurements in sky categorization. In terms of RMSD and MAD, the highest increase is seen in clear-sky and overcast categories. The MBD is positive in clear-sky category and negative in intermediate-cloudiness and overcast categories, which again shows that there are less amount of clouds in the clear-sky category and more clouds in intermediate-cloudiness and overcast categories. From a solar energy-harvesting point of view, the clear-sky days produce more energy than intermediate-cloudy or overcast days. It can be observed that ground-measurement and ERA5 predicts almost the same percentage of clear-sky days, which further shows that on daily averages reanalyses may not predict clouds accurately but on longer time scales, the solar radiation estimation improves.

Table 5

RMSD, MAD, and MBD for ERA5 daily averages in different sky categories. The errors are shown for the days when ERA5 and ground measurements agree on classification and for when they do not agree. Years from 2000 to 2015 are used in the analysis over all locations included in the study (see appendix B).

	Agreement on sky conditions			Disagree on sky conditions		
	RMSD (Wm^{-2})	MAD (Wm^{-2})	MBD (Wm^{-2})	RMSD (Wm^{-2})	MAD (Wm^{-2})	MBD (Wm^{-2})
Clear-sky	16.9	11.8	5.6	42.9	31.2	31.2
Intermediate-cloudiness	25.7	17.7	-7.4	33.8	24.2	-15.1
Overcast	15.3	9.6	-4.5	38.4	26.3	-26.3

5. Conclusion

This study provides a comprehensive evaluation of different GHI estimating datasets for high-latitude locations. Overall, SARAH provides lower errors than other datasets but is limited to 60-65ºN latitudes in Scandinavia; hence, it cannot provide complete coverage on the northern Scandinavian locations. For monthly averages of GHI, MAD of 5.8 Wm^{-2} is found for SARAH. Nevertheless, it provides very high quality solar-radiation estimates for the area it covers. The second best dataset found in this study is CLARA that has a global coverage and provides multi-decadal time series. For monthly mean estimates of GHI, CLARA gives a MAD of 6.3 Wm^{-2}. One of the challenges for estimating GHI at high latitude locations is the ability of the satellite estimation algorithms to differentiate between clouds and snow covers. ERA5 being a coarse-resolution global dataset is observed to perform nearly as well as CLARA with a MAD of 6.4 Wm^{-2} for monthly averages of GHI. ERA5 has similar spatial resolution as CLARA but it provides data on higher temporal resolutions and unlike CLARA, it has no missing values. ASR is found to have the highest errors in this analysis. MAD of 14.5 Wm^{-2} is found for ASR monthly means. In a similar study performed by Urraca et al. (2017b), MAD of 8 – 13 Wm^{-2} was reported for CM-SAF daily means datasets.

Both satellite estimation and reanalyses have problems in estimating solar radiation in intermediate-cloudiness and overcast conditions. To evaluate the strength of the datasets, the ground-measured data is divided into clear-sky, intermediate-cloudiness, and overcast categories and error statistics are calculated. In this test, satellite based estimations perform better than reanalyses. However, ERA5 has larger errors than CLARA and SARAH, but still considerably smaller errors than ASR. At high latitude locations, the seasonal variation in the length of the day is extreme. Taking this into consideration, an analysis is performed for different seasons. In this analysis, CLARA, SARAH and ERA5 have similar errors in the range of 6-13 Wm^{-2} in the summer months; however, ASR has relatively high errors in all seasons. On yearly GHI averages, SARAH provides the lowest MAD of 3.9 Wm^{-2}, followed by 4.8 Wm^{-2} for CLARA, 5.6 Wm^{-2} for ERA5, and 17.8 Wm^{-2} for ASR. SARAH and CLARA also provide better yearly energy estimates than ERA5 and ASR. CLARA and ERA5 are observed to provide lower errors below 65ºN than above, while CLARA and SARAH perform better at coastal regions, and ERA5 performs better in inland locations that have more snow covers.

Finally, an in-depth analysis is performed on ERA5 for its compatibility in sky stratification. It is found that in clear-sky conditions, the TCWC is overestimated, while in intermediate-cloudiness and overcast conditions it is underestimated. It is also observed that ERA5 has a positive bias on estimating clear-sky and intermediate-cloudy conditions, while a negative bias is seen in estimating overcast conditions. In conclusion, both CLARA and SARAH provide good estimates but both of these datasets have disadvantages, including the spatial limits of SARAH and the low temporal frequency of CLARA. On the other hand, ERA5 provides advantages in the form of historical data series and global coverage. On the basis of these results it is suggested that CLARA and SARAH provides better estimates for solar radiation, but ERA5 can be used to fill the missing data in these datasets.

Acknowledgements

We would like to thank NIBIO, ECMWF and CM-SAF for providing the data used in this study. This work is supported by Troms county and industry development fund under the project title, "Renewable energy in the arctic - academy and business in a joint effort" RDA12/46. Data was partly processed at the Stallo supercomputer at the University of Tromsø (UiT) provided by Norwegian Metacenter for Computational Science (NOTUR), project no.: NN9348k.

Appendix A

Table A1

Locations of the Norwegian measurement stations analysed in this study.

	Station	Latitude	Longitude	Altitude	Land type
1	Holt	69.65	18.91	12	Coastal
2	Sortland	68.65	15.28	14	Coastal
3	Vågønes	67.28	14.45	26	Coastal
4	Tjøtta	65.83	12.43	10	Coastal
5	Skogmo	64.51	12.02	32	Inland
6	Rissa	63.59	9.97	23	Coastal
7	Kvithamar	63.49	10.88	28	Inland
8	Skjetlein	63.34	10.3	44	Coastal
9	Surnadal	62.98	8.69	5	Inland
10	Tingvoll	62.91	8.19	23	Coastal
11	Fåvgang	61.46	10.19	184	Inland
12	Fureneset	61.29	5.04	12	Coastal
13	Gausdal	61.22	10.26	375	Inland
14	Løken	61.12	9.06	527	Inland
15	Ilseng	60.8	11.2	182	Inland
16	Kise	60.77	10.81	129	Inland
17	Apelsvoll	60.7	10.87	262	Inland
18	Hønefoss	60.14	10.27	126	Inland
19	Årnes	60.13	11.39	162	Inland
20	Etne	59.66	5.95	8	Inland
21	Ås	59.66	10.78	94	Inland
22	Bø	59.42	9.03	105	Inland
23	Rakkestad	59.39	11.39	102	Inland
24	Ramnes	59.38	10.24	39	Coastal
25	Tomb	59.32	10.81	12	Coastal
26	Gjerpen	59.23	9.58	41	Coastal
27	Hjelmeland	59.23	6.15	43	Inland
28	Tjølling	59.05	10.13	19	Coastal
29	Særheim	58.76	5.65	90	Coastal
30	Landvik	58.34	8.52	10	Coastal
31	Lyngdal	58.13	7.05	4	Inland

548 **Appendix B**

549 **Table B1**

550 List of years not included in the study.

	Station	Years having more than 5% missing data	Years failing Long and Dutton test	Years having operational error (snow/frost/ shading/soiling)	Years having equipment error
1	Holt	2001,2002,2006,2007,2008,2010	2013		2000
2	Sortland	2000,2006,2007,2010,2013			
3	Vågønes	2006,2007		2002	
4	Tjøtta	2006,2007			2008, 2012
5	Skogmo	2006,2007,2008,2015		2011	2013, 2014
6	Rissa	2006,2007	2000		
7	Kvithamar	2006,2007,2013			
8	Skjetlein	2006,2007	2000		
9	Surnadal	2006,2007,2014			
10	Tingvoll	2006,2007,2012			
11	Fåvang	2006,2007			2001
12	Fureneset	2006,2007,2011,2012			
13	Gausdal	2006,2007,2009			2015
14	Løken	2006,2007			
15	Ilseng	2006,2007,2004	2000	2009	
16	Kise	2002,2006,2007,2015		2013	
17	Apelsvoll	2006,2007		2002,2003,2004	2009
18	Hønefoss	2006,2007	2000		
19	Årnes	2006,2007			
20	Etne	2006,2007		2004,2012	
21	Ås	2006,2007			
22	Bø	2000,2006,2007			
23	Rakkestad	2006,2007			
24	Ramnes	2006,2007		2009	
25	Tomb	2006,2007	2009		
26	Gjerpen	2006,2007,2015			
27	Hjelmeland	2006,2007			2002, 2015
28	Tjølling	2006,2007,2008,2014		2012,2015	2009, 2010
29	Særheim	2000,2006,2007			
30	Landvik	2006,2007		2005,2010,2014, 2015	
31	Lyngdal	2006,2007	2001		

551

552

Appendix C

Table C1

Error metrics expressed in Wm^{-2}, for the datasets analysed in this study. Number without parentheses are monthly averaged errors while in parentheses are daily averaged errors.

Station	RMSD(Wm^{-2})				MAD(Wm^{-2})				MBD(Wm^{-2})			
	CLARA	SARAH	ERA5	ASR	CLARA	SARAH	ERA5	ASR	CLARA	SARAH	ERA5	ASR
Holt	5.5 (9.1)	-	4.7 (12.0)	9.6 (18.8)	1.5 (2.9)	-	1.1 (3.5)	2.8 (4.8)	-1.5 (-1.4)	-	1.1 (1.4)	2.5 (3.4)
Sortland	17.5 (23.0)	-	12.5 (29.9)	15.1 (38.2)	11.4 (16.0)	-	7.7 (18.9)	9.7 (24.4)	-11.0 (-12.0)	-	1.1 (2.4)	0.6 (2.2)
Vågønes	5.1 (13.8)	-	10.4 (26.7)	20.9 (42.1)	3.2 (8.7)	-	5.7 (15.0)	12.8 (24.8)	-0.7 (0.3)	-	3.9 (6.3)	11.8 (16.2)
Tjøtta	6.1 (13.8)	-	12.8 (29.2)	27.8 (47.9)	4.6 (9.5)	-	8.4 (17.4)	16.9 (28.0)	-0.3 (1.3)	-	7.9 (10.7)	15.2 (19.4)
Skogmo	12.4 (20.0)	11.8 (20.8)	8.2 (23.6)	20.2 (41.6)	7.8 (13.3)	8.2 (13.4)	5.3 (14.2)	12.5 (25.8)	-3.7 (-2.4)	-6.3 (-5.6)	1.0 (2.5)	11.4 (15.3)
Rissa	8.2 (17.3)	7.2 (17.6)	8.2 (27.1)	24.1 (45.4)	5.5 (12.3)	4.9 (11.5)	5.1 (17.1)	14.8 (27.7)	-2.7 (-1.5)	-3.3 (-2.4)	2.1 (4.2)	13.6 (19.5)
Kvithamar	7.3 (16.0)	7.8 (16.8)	7.7 (26.4)	31.6 (47.7)	5.1 (11.4)	1.0 (10.6)	5.2 (16.4)	20.3 (29.5)	-2.4 (-1.2)	-0.2 (-4.4)	-0.1 (1.4)	19.2 (23.1)
Skjetlein	7.9 (17.4)	8.8 (17.6)	7.2 (25.9)	29.9 (46.5)	6.0 (12.7)	6.4 (11.6)	5.1 (16.5)	19.7 (28.6)	-1.0 (0.8)	-6.0 (-4.8)	0.4 (2.0)	18.9 (22.6)
Surnadal	9.7 (20.8)	11.1 (23.5)	10.9 (28.4)	19.0 (41.2)	7.0 (14.1)	7.7 (14.5)	7.5 (17.9)	12.9 (25.2)	-4.1 (-2.7)	-6.0 (-5.1)	6.5 (8.3)	11.9 (14.8)
Tingvoll	8.3 (18.0)	9.3 (20.0)	10.4 (27.1)	16.9 (40.0)	6.4 (13.4)	6.4 (12.6)	6.5 (16.9)	10.7 (24.5)	-1.7 (-0.1)	-4.8 (-4.0)	5.1 (7.0)	8.4 (11.1)
Fåvang	13.1 (22.3)	10.0 (18.8)	10.4 (27.3)	21.2 (43.9)	9.5 (16.3)	7.3 (12.8)	6.8 (18.0)	14.6 (29.2)	-8.4 (-7.6)	-6.8 (-6.6)	1.8 (2.8)	14.1 (19.5)
Fureneset	4.7 (14.9)	5.7 (16.8)	10.7 (28.3)	18.5 (42.0)	3.5 (10.4)	3.8 (9.2)	6.9 (17.0)	12.6 (26.3)	-1.1 (0.6)	-2.8 (-1.8)	6.4 (8.6)	11.3 (14.1)
Gausdal	11.4 (20.6)	7.0 (17.4)	13.2 (27.7)	20.9 (42.6)	8.8 (15.3)	5.2 (12.1)	8.8 (18.3)	14.8 (29.0)	-1.2 (0.4)	-1.9 (-0.6)	5.1 (7.1)	14.6 (17.9)
Løken	14.3 (24.3)	10.7 (21.0)	9.2 (28.3)	12.7 (40.5)	9.5 (17.6)	7.4 (14.5)	5.9 (18.6)	8.2 (26.6)	-8.4 (-7.2)	-6.7 (-6.1)	1.8 (4.0)	3.7 (6.1)
Ilseng	11.8 (23.4)	9.3 (19.2)	11.0 (28.0)	21.4 (43.9)	8.8 (16.8)	5.8 (12.9)	7.6 (18.7)	16.7 (29.7)	-5.9 (-4.1)	-2.2 (-1.0)	-1.1 (1.0)	16.7 (19.8)
Kise	9.9 (20.6)	8.3 (18.2)	8.9 (25.9)	22.6 (42.7)	6.8 (15.1)	5.5 (12.4)	6.0 (17.0)	16.1 (28.6)	-1.0 (0.9)	-1.5 (0.4)	2.3 (4.3)	16.1 (20.8)
Apelsvoll	10.2 (19.6)	8.4 (17.4)	9.1 (25.8)	31.8 (48.2)	7.7 (14.7)	5.0 (11.9)	6.3 (17.3)	25.0 (34.0)	1.6 (4.2)	-0.2 (1.5)	2.8 (5.1)	25.0 (29.0)
Hønefoss	7.0 (16.6)	7.1 (15.8)	8.9 (25.7)	20.8 (41.6)	5.3 (12.2)	4.8 (10.8)	6.1 (17.0)	15.2 (28.0)	-3.3 (-1.3)	-3.5 (-2.2)	-0.1 (2.0)	15.1 (20.3)
Årnes	9.0 (17.3)	7.8 (16.2)	7.9 (24.6)	19.8 (40.0)	6.2 (12.7)	5.1 (11.0)	5.2 (16.3)	14.9 (26.6)	-4.0 (-2.8)	-3.8 (-3.0)	-1.3 (0.1)	13.3 (18.4)
Etne	9.3 (20.0)	9.7 (22.0)	12.8 (29.4)	23.1 (48.6)	6.9 (14.5)	7.0 (14.8)	8.9 (19.6)	15.0 (31.0)	-4.4 (-2.9)	-5.4 (-4.6)	6.4 (8.9)	14.5 (19.4)
Ås	7.3 (13.6)	7.1 (14.6)	8.0 (24.5)	21.1 (41.0)	4.8 (8.7)	5.1 (10.0)	5.3 (16.1)	15.1 (26.7)	-3.5 (-1.9)	-4.0 (-2.8)	-2.1 (-0.5)	14.4 (19.2)
Bø	7.9 (17.8)	6.5 (16.6)	10.3 (25.4)	21.5 (43.2)	5.7 (13.0)	4.6 (11.6)	7.1 (17.1)	16.4 (29.4)	1.4 (3.0)	1.5 (3.0)	4.9 7.5)	16.2 (20.3)
Rakkestad	7.2 (15.9)	7.8 (17.8)	8.2 (26.1)	21.0 (40.6)	5.5 (11.5)	5.5 (10.3)	5.8 (16.4)	16.1 (27.2)	-2.9 (-1.4)	-4.3 (-3.6)	0.6 (2.5)	15.2 (18.3)
Ramnes	8.9 (16.7)	7.5 (15.4)	8.2 (24.0)	22.1 (40.8)	7.1 (12.3)	5.5 (10.6)	5.8 (15.8)	16.4 (26.5)	-5.6 (-4.1)	-3.4 (-2.0)	-1.3 (0.3)	15.8 (18.7)
Tomb	11.5 (19.0)	12.7 (19.0)	11.2 (28.0)	20.3 (40.9)	7.0 (12.8)	8.9 (14.0)	6.9 (17.7)	14.4 (25.6)	-5.9 (-4.5)	-5.0 (-3.9)	-3.3 (-2.1)	12.5 (16.9)
Gjerpen	11.5 (19.1)	8.8 (19.9)	11.6 (25.8)	20.4 (40.8)	8.3 (14.5)	1.1 (14.3)	8.4 (18.1)	14.8 (27.9)	-4.3 (-3.0)	-0.2 (-4.1)	-1.3 (0.3)	10.7 (15.4)
Hjelmeland	4.7 (16.6)	5.5 (16.1)	10.9 (29.5)	19.7 (46.0)	3.4 (12.1)	3.6 (10.9)	7.5 (19.6)	13.9 (31.1)	-0.1 (1.6)	-0.9 (0.2)	6.6 (9.8)	13.5 (18.1)
Tjølling	8.2 (18.0)	7.5 (13.8)	11.4 (26.5)	28.5 (45.1)	6.0 (12.8)	5.2 (9.6)	7.5 (16.8)	19.7 (29.6)	-0.1 (1.9)	-1.9 (-0.6)	4.4 (7.2)	19.1 (25.1)
Særheim	5.9 (15.2)	6.2 (16.0)	7.4 (26.4)	17.0 (43.5)	4.4 (10.8)	4.3 (10.8)	4.9 (16.7)	11.6 (28.0)	-1.2 (0.3)	-1.9 (-0.7)	2.2 (4.1)	9.7 (12.8)
Landvik	7.3 (16.5)	6.3 (14.3)	10.2 (25.7)	23.0 (42.2)	5.2 (11.6)	4.6 (9.7)	6.5 (16.7)	15.8 (28.1)	1.2 (3.9)	-0.4 (1.8)	5.5 (8.8)	15.5 (21.3)
Lyngdal	11.3 (21.8)	9.6 (22.8)	12.1 (29.9)	15.8 (41.5)	7.4 (13.7)	1.2 (13.5)	8.3 (19.5)	10.9 (25.9)	-6.0 (-6.1)	-0.4 (-1.8)	2.9 (3.6)	5.9 (7.0)

APPENDIX D

Table D1

Statistical errors of the yearly average solar radiation for the datasets included in the study. This table shows the deviations for inland, coastal, above 65 ° N, and below 65ºN latitude regions. RMSD, MAD, and MBD are expressed in Wm^{-2}.

	RMSD ($Wm^{-2}year^{-1}$)				MAD ($Wm^{-2}year^{-1}$)				MBD ($Wm^{-2}year^{-1}$)			
	CLARA	SARAH	ERA5	ASR	CLARA	SARAH	ERA5	ASR	CLARA	SARAH	ERA5	ASR
All Sites	7.4	5.2	6.8	18.7	4.8	3.9	5.6	17.8	-4.2	-2.8	4.4	17.5
Above 65ºN	8.9	-	9.6	16.6	5.2	-	8.8	15.6	-4.4	-	7.1	13.4
Below 65ºN	7.2	5.2	6.4	18.9	4.8	3.9	5.2	18.1	-4.1	-2.9	4.1	18.0
Coastal	6.2	4.9	7.4	18.1	3.8	3.5	6.2	17.1	-3.1	-2.4	4.6	16.6
Inland	8.2	5.4	6.4	19.2	5.6	4.0	5.0	18.4	-5.0	-3.0	4.2	18.2

Appendix E

Table E1

Seasonal analysis of the datasets showing the variations in terms of RMSD, MAD, and MBD and expressed in Wm^{-2}. CLARA and SARAH performs similarly and better than other datasets, while ERA5 gives median values and ASR performs the worst

	RMSD (Wm^{-2})				MAD (Wm^{-2})				MBD (Wm^{-2})			
	FMA	MJJ	ASO	NDJ	FMA	MJJ	ASO	NDJ	FMA	MJJ	ASO	NDJ
CLARA	12.7 (29.0)	9.4 (22.3)	8.4 (16.6)	1.9 (13.1)	8.8 (17.3)	7.3 (17.0)	6.4 (11.5)	0.5 (7.9)	-7.6 (-11.7)	1.1 (1.3)	-4.1 (0.4)	-0.3 (-7.1)
SARAH	12.3 (21.9)	9.8 (25.5)	8.8 (15.9)	5.5 (10.2)	9.2 (15.0)	7.8 (17.9)	6.8 (11.0)	3.5 (4.8)	-7.4 (-7.2)	-2.8 (-3.1)	-4.0 (1.2)	-2.9 (-2.4)
ERA5	11.5 (23.8)	15.1 (41.3)	9.9 (24.1)	4.7 (9.7)	8.8 (17.0)	12.2 (31.8)	7.5 (17.0)	2.7 (4.5)	6.9 (7.3)	7.3 (7.3)	-3.1 (2.2)	-0.1 (0.4)
ASR	22.5 (35.9)	38.4 (67.6)	17.0 (38.6)	5.5 (11.3)	19.4 (26.0)	34.6 (52.8)	12.8 (27.1)	3.4 (5.8)	18.9 (19.9)	34.0 (34.2)	9.2 (14.3)	1.2 (1.5)

References

Babar, B., Graversen, R., Boström, T., 2018. Evaluating CM-SAF solar radiation CLARA-A1 and CLARA-A2 datasets in Scandinavia. Solar Energy 170, 76-85.

Bird, R.E., Hulstrom, R.L., 1981. Simplified clear sky model for direct and diffuse insolation on horizontal surfaces. Solar Energy Research Inst., Golden, CO (USA).

Boilley, A., Wald, L., 2015. Comparison between meteorological re-analyses from ERA-Interim and MERRA and measurements of daily solar irradiation at surface. Renewable Energy 75, 135-143.

Bojanowski, J.S., Vrieling, A., Skidmore, A.K., 2014. A comparison of data sources for creating a long-term time series of daily gridded solar radiation for Europe. Solar Energy 99, 152-171.

Bromwich, D., Wilson, A., Bai, L., Liu, Z., Barlage, M., Shih, C.-F., Maldonado, S., Hines, K., Wang, S.-H., Woollen, J., 2017. The Arctic System Reanalysis Version 2. Bulletin of the American Meteorological Society(2017).

Cano, D., Monget, J.-M., Albuisson, M., Guillard, H., Regas, N., Wald, L., 1986. A method for the determination of the global solar radiation from meteorological satellites data. Solar energy 37(1), 31-39.

Crabtree, G., Misewich, J., Ambrosio, R., Clay, K., DeMartini, P., James, R., Lauby, M., Mohta, V., Moura, J., Sauer, P., 2011. Integrating renewable electricity on the grid, AIP Conference Proceedings. AIP, pp. 387-405.

Dee, D.P., Uppala, S.M., Simmons, A., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., Bauer, d.P., 2011. The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. Quarterly Journal of the royal meteorological society 137(656), 553-597.

ECMWF, 2018. ERA5 data documentation.

Gautier, C., Diak, G., Masse, S., 1980. A simple physical model to estimate incident solar radiation at the surface from GOES satellite data. Journal of Applied meteorology 19(8), 1005-1012.

Gueymard, C.A., 2011. Uncertainties in modeled direct irradiance around the sahara as affected by aerosols: Are current datasets of bankable quality? Journal of Solar Energy Engineering 133(3), 031024.
Hall, I.J., Prairie, R., Anderson, H., Boes, E., 1978. Generation of a typical meteorological year. Sandia Labs., Albuquerque, NM (USA).
Hammer, A., Heinemann, D., Hoyer, C., Kuhlemann, R., Lorenz, E., Müller, R., Beyer, H.G., 2003. Solar energy assessment using remote sensing technologies. Remote Sensing of Environment 86(3), 423-432.
Hans, H., Dick, D., 2016. ERA5 reanalysis is in production. https://www.ecmwf.int/en/newsletter/147/news/era5-reanalysis-production.
Heinemann, D., Lorenz, E., Girodo, M., 2006. Forecasting of solar radiation. Solar energy resource management for electricity generation from local level to global scale. Nova Science Publishers, New York.
Huld, T., Paietta, E., Zangheri, P., Pinedo Pascua, I., 2018. Assembling Typical Meteorological Year Data Sets for Building Energy Performance Using Reanalysis and Satellite-Based Data. Atmosphere 9(2), 53.
Ineichen, P., 2014. Long term satellite global, beam and diffuse irradiance validation. Energy Procedia 48, 1586-1596.
Juruš, P., Eben, K., Resler, J., Krč, P., Kasanický, I., Pelikán, E., Brabec, M., Hošek, J., 2013. Estimating climatological variability of solar energy production. Solar Energy 98, 255-264.
Karlsson, K.-G., Anttila, K., Trentmann, J., Stengel, M., Meirink, J.F., Devasthale, A., Hanschmann, T., Kothe, S., Jääskeläinen, E., Sedlar, J., 2017. CLARA-A2: the second edition of the CM SAF cloud and radiation data record from 34 years of global AVHRR data. Atmospheric Chemistry and Physics 17(9), 5809.
Kennedy, A.D., Dong, X., Xi, B., Xie, S., Zhang, Y., Chen, J., 2011. A comparison of MERRA and NARR reanalyses with the DOE ARM SGP data. Journal of Climate 24(17), 4541-4557.
Kishore, P., Ratnam, M.V., Namboothiri, S., Velicogna, I., Basha, G., Jiang, J., Igarashi, K., Rao, S., Sivakumar, V., 2011. Global (50 S–50 N) distribution of water vapor observed by COSMIC GPS RO: Comparison with GPS radiosonde, NCEP, ERA-Interim, and JRA-25 reanalysis data sets. Journal of Atmospheric and Solar-Terrestrial Physics 73(13), 1849-1860.
Long, C.N., Dutton, E.G., 2010. BSRN Global Network recommended QC tests, V2. x.
Mueller, R., Behrendt, T., Hammer, A., Kemper, A., 2012. A new algorithm for the satellite-based retrieval of solar surface irradiance in spectral bands. Remote Sensing 4(3), 622-647.
Mueller, R., Träger-Chatterjee, C., 2014. Brief accuracy assessment of aerosol climatologies for the retrieval of solar surface radiation. Atmosphere 5(4), 959-972.
Müller, R., Pfeifroth, U., Träger-Chatterjee, C., Trentmann, J., Cremer, R., 2015. Digging the METEOSAT treasure—3 decades of solar surface radiation. Remote Sensing 7(6), 8067-8101.
Noia, M., Ratto, C., Festa, R., 1993. Solar irradiance estimation from geostationary satellite data: II. Physical models. Solar Energy 51(6), 457-465.
Pfeifroth, U., Kothe, S., Müller, R., Trentmann, J., Hollmann, R., Fuchs, P., Werscheck, M., 2017a. Surface Radiation Data Set-Heliosat (SARAH)—Edition 2, Satellite Application Facility on Climate Monitoring. .
Pfeifroth, U., Kothe, S., Müller, R., Trentmann, J., Hollmann, R., Fuchs, P., Werscheck, M., 2017b. Surface Radiation Data Set–Heliosat (SARAH)–Edition 2, Satellite Application Facility on Climate Monitoring.
Pfeifroth, U., Sanchez-Lorenzo, A., Manara, V., Trentmann, J., Hollmann, R., 2018. Trends and Variability of Surface Solar Radiation in Europe Based On Surface- and Satellite-Based Data Records. Journal of Geophysical Research: Atmospheres, n/a-n/a.
Pinker, R., Laszlo, I., 1992. Modeling surface solar irradiance for satellite applications on a global scale. Journal of Applied Meteorology 31(2), 194-211.

Platt, C., 1983. On the bispectral method for cloud parameter determination from satellite VISSR data: Separating broken cloud and semitransparent cloud. Journal of climate and applied meteorology 22(3), 429-439.
Polo, J., Antonanzas-Torres, F., Vindel, J., Ramirez, L., 2014. Sensitivity of satellite-based methods for deriving solar radiation to different choice of aerosol input and models. Renewable energy 68, 785-792.
Posselt, R., Mueller, R., Stöckli, R., Trentmann, J., 2012. Remote sensing of solar surface radiation for climate monitoring—The CM-SAF retrieval in international comparison. Remote Sensing of Environment 118, 186-198.
Remund, J., Perez, R., Lorenz, E., 2008. Comparison of solar radiation forecasts for the USA, Proc. of the 23rd European PV Conference. Valencia, Spain.
Rigollier, C., Lefèvre, M., Wald, L., 2004. The method Heliosat-2 for deriving shortwave solar radiation from satellite images. Solar Energy 77(2), 159-169.
Riihelä, A., Carlund, T., Trentmann, J., Müller, R., Lindfors, A.V., 2015. Validation of CM SAF surface solar radiation datasets over Finland and Sweden. Remote Sensing 7(6), 6663-6682.
Smith, C.J., Bright, J.M., Crook, R., 2017. Cloud cover effect of clear-sky index distributions and differences between human and automatic cloud observations. Solar Energy 144, 10-21.
Stoffel, T., Renne, D., Myers, D., Wilcox, S., Sengupta, M., George, R., Turchi, C., 2010. Concentrating solar power. Golden: National Renewable Energy Laboratory.
Suri, M., Cebecauer, T., 2014. Satellite-based solar resource data: Model validation statistics versus user's uncertainty, ASES SOLAR 2014 Conference, San Francisco. pp. 7-9.
Tarpley, J., 1979. Estimating incident solar radiation at the surface from geostationary satellite data. Journal of Applied Meteorology 18(9), 1172-1181.
Trentmann, J., Kothe, S., 2016. Surface Radiation Products

Product User Manual. CM SAF Cloud, Albedo, Radiation dataset,

SAF/CM/DWD/PUM/GAC/RAD(2.1).
Urraca, R., Gracia-Amillo, A.M., Huld, T., Martinez-de-Pison, F.J., Trentmann, J., Lindfors, A.V., Riihelä, A., Sanz-Garcia, A., 2017a. Quality control of global solar radiation data with satellite-based products. Solar Energy 158, 49-62.
Urraca, R., Gracia-Amillo, A.M., Koubli, E., Huld, T., Trentmann, J., Riihelä, A., Lindfors, A.V., Palmer, D., Gottschalg, R., Antonanzas-Torres, F., 2017b. Extensive validation of CM SAF surface radiation products over Europe. Remote sensing of environment 199, 171-186.
Urraca, R., Huld, T., Gracia-Amillo, A., Martinez-de-Pison, F.J., Kaspar, F., Sanz-Garcia, A., 2018. Evaluation of global horizontal irradiance estimates from ERA5 and COSMO-REA6 reanalyses using ground and satellite-based data. Solar Energy 164, 339-354.
Widén, J., Shepero, M., Munkhammar, J., 2017. On the properties of aggregate clear-sky index distributions and an improved model for spatially correlated instantaneous solar irradiance. Solar Energy 157, 566-580.
Wild, M., 2008. Short-wave and long-wave surface radiation budgets in GCMs: A review based on the IPCC-AR4/CMIP3 models. Tellus A 60(5), 932-945.
Wild, M., Folini, D., Henschel, F., Fischer, N., Müller, B., 2015. Projections of long-term changes in solar radiation based on CMIP5 climate models and their influence on energy yields of photovoltaic systems. Solar Energy 116, 12-24.
Yi, Y., Kimball, J.S., Jones, L.A., Reichle, R.H., McDonald, K.C., 2011. Evaluation of MERRA land surface estimates in preparation for the soil moisture active passive mission. Journal of Climate 24(15), 3797-3816.
You, Q., Sanchez-Lorenzo, A., Wild, M., Folini, D., Fraedrich, K., Ren, G., Kang, S., 2013. Decadal variation of surface solar radiation in the Tibetan Plateau from observations, reanalysis and model simulations. Climate dynamics 40(7-8), 2073-2086.

695 Zhao, L., Lee, X., Liu, S., 2013. Correcting surface solar radiation of two data assimilation systems
696 against FLUXNET observations in North America. Journal of Geophysical Research: Atmospheres
697 118(17), 9552-9564.
698 Øyvind, B., Anne, L., Løvholm, Sonia, L., 2013. Resource mapping of solar energy an overview of
699 available data in Norway. Kjeller Vindteknikk

700 Report KVT/OB/2013/R046.

701

Paper IV [redacted]

Manuscript prepared for Solar Energy ISSN: 0038-092X

Random forest regression for improved mapping of solar power resources at high latitudes

Bilal Babar*, Luigi Tommaso Luppino, Tobias Boström and Stian Normann Anfinsen

Department of Physics and Technology, The Arctic University – University of Tromsø, Norway

*corresponding author: bilal.babar@uit.no

Abstract

Datasets from meteorological reanalyses and measurements from polar orbiting satellites are the available sources of large-scale information about solar radiation. However, both the reanalyses and the satellite-based estimates can be severely biased, especially in high latitude regions. In this study, solar radiation estimates from the ECMWF Reanalysis 5 (ERA5) and the Cloud, Albedo, Radiation dataset Edition 2 (CLARA-A2) were used as input to a random forest regression (RFR) model to construct a novel dataset with higher accuracy and precision than the input datasets. For monthly averages of global horizontal irradiance (GHI) at Norwegian sites, CLARA-A2 and ERA5 respectively produced a root mean squared deviation (RMSD) of 9.6 Wm^{-2} and 10.2 Wm^{-2}, a mean absolute deviation (MAD) of 6.3 Wm^{-2} and 7.0 Wm^{-2}, and a bias of -1.6 Wm^{-2} and 3.9 Wm^{-2}. In contrast, the proposed regression model provided an RMSD of 6.6 Wm^{-2}, an MAD of 4.3 Wm^{-2}, and a bias of -0.2 Wm^{-2}. This shows that the RFR model is both accurate and precise, and significantly reduces both dispersion and bias in the new dataset with respect to the constituent sources. The proposed model provided more accurate and precise estimates in a seasonal error analysis as well. A sky stratification analysis was performed to evaluate the accuracy of the datasets under different sky conditions. It was found that the proposed model provides better estimates under all sky conditions with particular improvements in overcast conditions. The proposed regression model was also tested on five Swedish locations and it was found to improve solar radiation estimates to a similar degree as for the Norwegian locations, thus proving its consistency under similar climatic conditions.

Keywords: Solar radiation; High latitudes; ERA5; CLARA; CMSAF; Random forest regression

1. Introduction

The bankability of solar power plants largely depends on the accuracy and precision of the solar radiation measurements or estimates, which are required at all stages of solar energy projects. Time series or temporal averages of solar radiation are obtained initially before a particular system can be simulated and its design criteria and performance are evaluated. In the case of flat plate collectors, such as photovoltaic (PV) and thermal, global horizontal irradiance (GHI) or global tilted irradiance (GTI) are used in the feasibility and planning phases. Additionally, long-term variability in solar radiation is used to quantify the solar resource and project worst-case scenarios of energy production in such systems. During operation, real-time data are typically required to verify the performance of the system and detect problems. In both cases, the required data can be obtained from measurement, modelling, or a combination of both (Sengupta et al., 2017; Urraca et al., 2017b).

High quality solar resource assessments make technology deployment possible by helping the decision makers to reduce the uncertainty in investment decisions. However, the assessments cannot rely exclusively on ground measurements of solar radiation, because these are usually not available at most locations in the world. Even though such measurements exist at some locations, they frequently contain missing or erroneous data that must be filled in by using modelled data or interpolation from nearby measurement stations. Lastly, the cost of maintaining local equipment is larger than operating a model, assuming that satellite data and the output of reanalyses are provided free of charge or at a reasonable cost. Although model data are not as accurate as ground measurements, they can be used as an alternative

(Stoffel et al., 2010). Nevertheless, quality ground measurements remain essential because they have low errors and can be used to validate models (Sengupta et al., 2017).

Geostationary satellites are widely used for estimating surface solar radiation at low and medium latitudes, where their measurements of top-of-atmosphere upwelling radiances and surface albedos are used to derive GHI at the surface (Cano et al., 1986; Pinker and Laszlo, 1992; Rigollier et al., 2004; Tarpley, 1979). These satellites are positioned over the equator at different longitudes in order to provide a global coverage between -60° and +60° in latitudes. For instance, the Meteosat first and second generation geostationary satellites provide coverage of most of continental Europe (Müller et al., 2015; Pfeifroth et al., 2017; Schmetz et al., 2002; Urraca et al., 2017b). However, estimates above 65°N are prohibited by the slant viewing angle that geostationary satellites experience when they point away from nadir *i.e.*, the vertical direction directly below the satellite (Schulz et al., 2009).

Above the critical latitudes that limit geostationary satellites, polar orbiting satellites can be used to estimate surface solar radiation (Karlsson et al., 2017). Polar orbiting satellites traverse the entire Earth and provide global coverage, but their accuracy decrease at high latitudes because of the large angles between the satellite sensor and the Sun. Another factor that decreases the accuracy at high latitude is the frequent snow cover, which the satellites sensors cannot differentiate from clouds in the visible spectrum. The temporal resolution of solar radiation estimated by polar orbiting satellites is lower than that of geostationary satellites, since the revisit time of the former is higher than the repeat time of image acquisitions used by geostationary satellites. Whereas the latter capture images at least every 15 minutes, the polar orbiting satellites sense a given location twice each day on the equator and about 14 times each day near the poles. The sensing frequency of polar orbiting satellites is best at high latitudes, since swath overlap increases towards the poles, where their orbits converge. The accuracy of solar radiation estimated from satellite data is lower than ground measurements, but the advantages include large spatial and temporal coverage (Noia et al., 1993). In another study it was observed that estimates from polar orbiting satellites provide reasonable accuracy, but estimates obtained over snow-covered surfaces result in high errors because it is difficult to differentiate clouds from snow in the visible spectrum of light (Babar et al., 2018a). For a list of known issues and uncertainty sources, refer to Suri and Cebecauer (2014).

In addition to satellite measurements, meteorological reanalyses also provide surface short-wave incoming radiation estimates (Wild, 2008; Wild et al., 2015). Reanalysis datasets are produced by data assimilation of historical observational data, aiming to obtain the initial state of selected parameters which best fits a numerical weather prediction (NWP) model to the available data (Kennedy et al., 2011). Reanalyses are not as accurate as satellite-based estimates, but they provide global coverage for multi-decadal time range (Babar et al., 2018b; Urraca et al., 2017b; Urraca et al., 2018).

Both the satellite-derived estimates and reanalyses have a certain degree of uncertainty, but proper identification and removal of errors can improve the results. Site adaptation refers to the improvements that can be obtained in satellite-derived or model-based solar irradiance by using short-term ground measurements to reduce the systematic bias in the original dataset. In Polo et al. (2016), the authors have provided a preliminary survey of available site adaptation techniques. Site adaptation can be physically based methods in which the atmospheric input data such as aerosol optical depth and vertically-integrated water column are adjusted to better match the ground based observations (Gueymard, 2012). Other such methods include the use of clear-sky models to adjust the atmospheric aerosol on clear sky days (Cebecauer and Šúri, 2012). The second type of site adaptation is based on statistical adjustment of meteorological observations, such as rain, wind and so forth. The linear statistical methods for bias removal is performed by first fitting a line to the observations and estimations. In the next step an x=y line is subtracted from all observations (Polo et al., 2015). This type of adjustment removes the systematic errors that exist due to the regional inconsistencies or from the radiative models. Moreover, non-parametric regression by using multiple input datasets has been performed by Davy et al. (2016) for Australia. In this study, the authors used generalized additive models with cubic smoothing splines

to improve accuracy. By including an NWP model-derived irradiance as input, they reduced the root mean square deviation by a few percent. In the study presented here, an approach similar to the site adaptation technique by Davy et al. (2016) is used.

This study presents a novel dataset that is obtained by using mainly the solar radiation estimates from ECMWF Reanalysis 5 (ERA5) and Cloud, Albedo, Radiation dataset Edition 2 (CLARA-A2), hereafter referred as ERA5 and CLARA. It is observed that reanalyses usually overestimate surface solar radiation and satellite methods usually underestimate it (Babar et al., 2018a; Riihelä et al., 2015; Urraca et al., 2017b). The main motivation behind constructing a new estimate is that we want to overcome the underestimation tendency of satellite methods and the overestimation tendency of reanalyses by combining them into a dataset with lower bias and variance. The input datasets were used together with *in-situ* measurements to develop a novel random forest regression (RFR) model, which can be used to produce accurate and precise estimates of solar radiation at high latitudes.

This paper is formatted as follows: Section 2 describes the datasets, quality control procedures, RFR model and pre-processing used in this study. Section 3 describes the results of the study. Section 4 provides a conclusion of this work.

2. Datasets

CLARA and ERA5 are coarse resolution datasets and provide data on a grid of 0.25° x 0.25° and 0.28° x 0.28°, respectively. Data extraction from these datasets is performed by selecting the four grid points surrounding any location where we have ground measurements, and applying inverse distance weighted interpolation to obtain solar radiation at these coordinates. In case of CLARA, there are missing data points, which implies that at some of the time frames there is data lacking in the surrounding four grid points. When the surrounding points have less than three valid values, the interpolation is replaced by a missing data value, indicating that a valid value could not be extracted for that particular time. The ERA5 dataset does not contain missing values. It will be explained in section 2.6 how the proposed regression model handles missing data values.

2.1 CLARA-A2

This dataset was released in December 2016 and it is the second edition of CLARA (Cloud, Albedo, Radiation dataset) produced by Eumetsat's Satellite Application Facility on Climate Monitoring (CM-SAF) (Karlsson et al., 2017). The dataset covers 1 January 1982 to 31 December 2015, and constitutes an extension of 6 years relative to the previous CLARA-A1 dataset. This dataset has global coverage with a spatial resolution of 0.25° x 0.25° on a regular latitude-longitude grid and it provides daily and monthly averages of surface incoming shortwave (SIS) radiation. To calculate daily averages, at least 20 observations of incoming solar radiation in each grid box are required. Similarly, 20 valid daily averages are required to generate monthly averages (SAF, 2016). Along with SIS, CLARA also provides longwave up- and down-welling surface radiation.

The fundamental method used in calculating surface solar irradiance from satellite observations is that the reflectance measured by the satellite instruments is related to the atmospheric transmittance. The underlying algorithm in CLARA uses Advanced Very High Resolution Radiometer (AVHRR) sensor data to derive the cloud cover, which is used to calculate surface incoming solar radiation (Karlsson et al., 2017). In addition to the cloud cover information, the solar radiation is estimated by using auxiliary data like the solar zenith angle, vertically-integrated water vapour and aerosol optical depth. Finding solar zenith angles is straightforward and can be calculated accurately. In this dataset, all data points with solar zenith angles larger than 80° are set to missing values and solar zenith angles larger than 90° are set to zero. The vertically-integrated water vapour and aerosol optical depth are not available in the AVHRR data and for these external sources are used. For vertically-integrated water vapour, the ERA-Interim Reanalysis (Dee et al., 2011) is used and the vertical ozone column is set to a constant value of 335 DU, as its variability has negligible impact on the estimated solar radiation. Aerosol information

for the algorithm is taken from the modified version of the monthly mean aerosol fields from the Global Aerosol Data Set/Optical Properties of Aerosols and Cloud (GADS/OPAC) climatology. In the algorithm, AVHRR data is used to retrieve only the cloud cover information. The first step in estimating surface solar radiation is the classification of the sky condition. Software from Eumetsat's Nowcasting Satellite Application Facility (SAFNWC) is used to derive the information on cloud coverage for each pixel by using the information from the satellite sensor (SAF, 2016). If no cloud is detected (cloud free pixel), surface solar radiation is calculated by using the clear-sky Mesoscale Atmospheric Global Irradiance Code (MAGIC) (Mueller et al., 2009) by using only auxiliary sources. If the pixel is classified as cloudy (cloud contaminated or fully cloudy), visible channels of AVHRR instrument are used to derive broadband reflectance. The reflectance for each pixel is then transferred to broadband fluxes by using a bidirectional reflectance distribution function (BRDF). In the next step, the broadband top-of-the-atmosphere albedo is used to derive transmissivity through a look-up table approach. Finally, the transmissivity is used to calculate the surface solar radiation. However, as a temporally constant surface albedo is used by the algorithm, it does not provide radiation estimates on snow and sea ice coverage areas (Karlsson et al., 2017). For more information on the CLARA dataset and its accuracy, refer to Karlsson et al. (2017).

2.2 ERA5

ECMWF Reanalysis 5 (ERA5) is the fifth generation atmospheric reanalysis of the global climate from the European Centre for Medium-Range Weather Forecasts (ECMWF). It spans a period from 1950 to near present time (Hersbach and Dee, 2016). At the time of this study, data from 2000 to 2017 is available. Further data back in time will be released in 2019-20, and the dataset will continue to update forward in near real-time. In ERA5, the solar radiation variable has a spatial resolution of 31km (0.28125° x 0.28125°) and an hourly temporal frequency. ERA5 uses Integrated Forecasting System (IFS) cycle 41r2 with a state-of-the-art four-dimensional variational analysis (4DVAR) assimilation system. ERA5 has a higher number of pressure levels than ERA-Interim (the previous edition of ECMWF reanalysis) and provides more parameters, including hourly estimates of atmospheric, land and oceanic climate variables. For more information on ERA5 refer to ECMWF (2018).

In this study, shortwave surface downward radiation and shortwave surface downward radiation clear-sky are used from this dataset. In ERA5, the incoming shortwave irradiance is obtained from a Radiative Transfer Model (RTM). This model simulates the attenuation in solar radiation caused by the atmosphere. Therefore, the quality of the radiation estimates depends on the RTM used. Reanalyses generally do not assimilate aerosol, clouds or water vapour data, which increases the uncertainty in the estimated surface irradiance (You et al., 2013; Zhao et al., 2013).

2.3 Ground data

The ground-measured data used in this study for regression and validation is obtained from the Norwegian Institute of Bioeconomy Research (NIBIO) for Norwegian locations and from the Swedish Meteorological and Hydrological Institute (SMHI) for Swedish locations. NIBIO and SMHI collect, maintain, and provide data from their respective networks of meteorological measurement stations in Norway and Sweden, including ground-measured solar radiation. NIBIO and SMHI register hourly-average GHI by using Kipp and Zonen CMP11 or CMP13 pyranometers. The data is quality controlled and the equipment is maintained regularly on a daily or weekly basis (NIBIO, 2018; Persson, 2000). The coordinates of the locations, their altitudes and land type are indicated in Appendix A, Tables A1-A2 and an overview of the site locations is shown in Figure 1. The Swedish locations were only used in the testing of regression model, so as to prove its robustness.

For the analysis, the Norwegian sites were divided into inland and coastal regions by observing the proximity to the shoreline. Regions within 30 km of the shoreline were considered as coastal. From the 31 Norwegian locations studied here, 14 sites were classified as coastal and 17 sites as inland. The

locations were also divided into two other groups, where locations lying above 65°N were grouped together and locations lying below 65°N were put in another group. In this latitude-based grouping, four sites were in the above 65°N group and 27 sites belonged to below 65°N group. For details on this classification, refer to appendix A, Table A1.

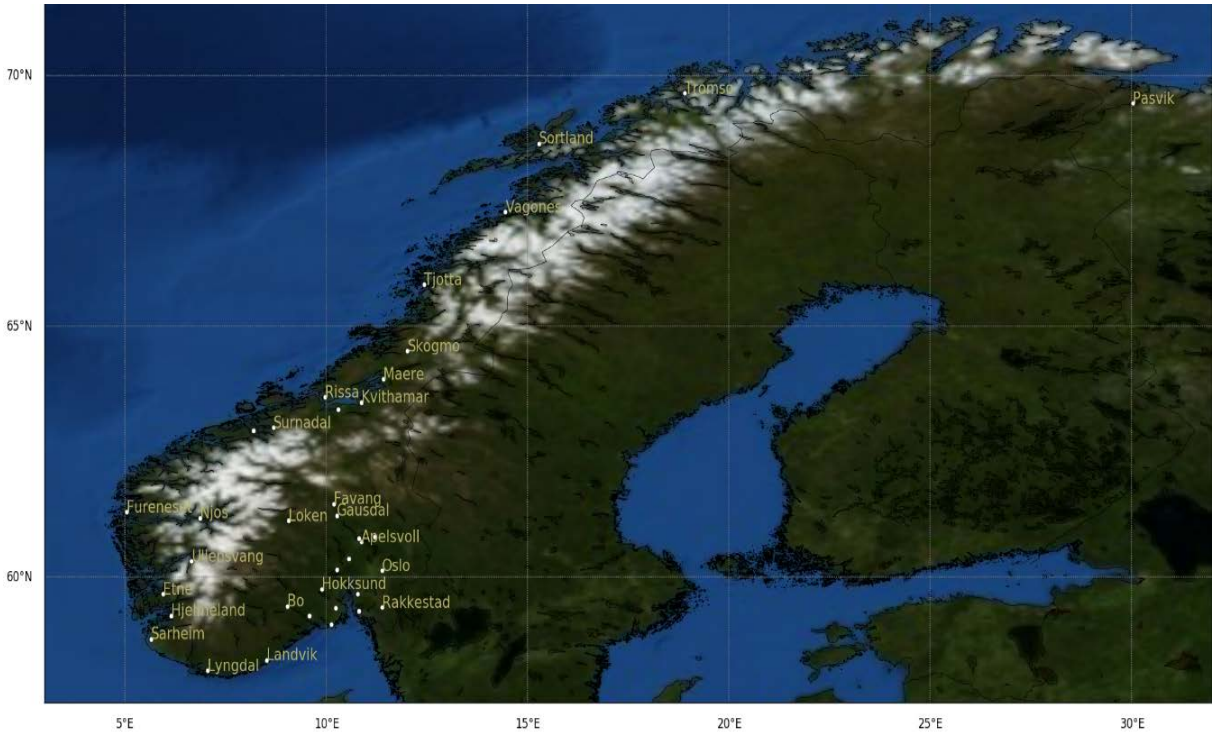


Figure 1: Locations of the Norwegian sites included in the study. To avoid overlapping of names some locations are shown with only white dots.

2.4 Quality Control

Although the data provided by NIBIO are quality controlled, Urraca et al. (2017a) observed that operational and equipment errors exist in NIBIO stations. The first check performed in this study is to look at the percentage of missing data. Any year having more than 5% of missing values was discarded from the analysis. The second check was performed by using the BSRN Global Network recommended quality control (QC) tests, version 2.0 (Long and Dutton, 2010). The BSRN QC test highlights values that are extremely rare and physically impossible. Based on this test, years having more than 1% of flagged values were removed from the ground data. The third quality control procedure was applied by using the QC technique of Urraca et al. (2017a). In this test, CLARA and ERA5 datasets are used to check the quality of ground measurements by constructing confidence intervals to detect the operational and equipment errors. Following Urraca et al. (2017a), the locations in Norway were divided into two sections by grouping locations above 65°N and locations below 65°N. Separate confidence intervals were constructed for both groups. After constructing these confidence intervals, the ground data was passed through an algorithm to check the data with errors, which appear in the form of flags. Following Urraca et al. (2017a) two checks were performed, one to see the operational errors and the other to see the equipment errors. After these checks, the years having large number of flags were visually inspected and removed from the analysis. For example, Pasvik, Mære, Njøs and Ullensvang were found to have a large number of flags from the third QC test, hence these locations were discarded. For more information on this quality control procedure, refer to Urraca et al. (2017a). A number of Norwegian locations were found to have large percentage of missing data points in years 2006 and 2007, hence these years were

rejected from all Norwegian locations. See Appendix B, Table B1 for details of the years not included in this study.

2.5 Random forest regression

The motivation for adopting a regression model from the recent machine learning literature came from the hypothesis that our regression analysis might benefit from using an algorithm, which applies different regression functions for different subsets of the predictor data space. Conventional regression methods apply the same regression function, parametric or nonparametric, to the whole dataset and include all independent variables (predictors) as arguments to this function. More advanced methods can, on the other hand, allow more flexibility by judiciously selecting subsets of predictors or tailoring the regression function for subsets of the data in a manner that improves the overall performance in the regression analysis.

An example of such an approach is stratified regression analysis (Anderson et al., 1980; Tso and Yau, 2007), where separate regression models are set up for stratified samples of the independent variables, that are observed or hypothesized to exhibit different relations to the dependent variable. The strata can often be identified directly from the independent variables as natural groupings of the data. This idea is further developed in so-called clusterwise regression or regression clustering (Bagirov et al., 2017; Hsu, 2015; Späth, 1979), where clusters in the independent data are identified during the adaption of the regression model. Both the cluster-specific regression functions and the optimal clustering of the independent variable space are learnt iteratively from a training dataset containing paired independent and dependent samples. Input data points (vectors of predictor data) may be assigned to a unique cluster, or they may be given a fuzzy membership in multiple clusters. These membership values may then be used as weights in an ensemble approach where the dependent variable is predicted as a weighted average of the clusterwise regression functions. Another approach is the use of regression trees (Tso and Yau, 2007; Yu et al., 2010), where the predictor data space is recursively partitioned into finer regions using a tree structure, hoping that stronger relationships between independent and dependent variables can be formulated in these fine regions or branches of the tree. This may capture relations that are difficult to perceive in an explanatory data analysis if structures in the data are not visually apparent.

RFR is a regression tree method that has become very popular in recent years due to its strong performance, ease of implementation and low computational cost. It is an ensemble learning technique developed by Leo Breiman (Breiman, 2001), which is based on the construction of a multitude of decision trees. Branches of the trees represent particular paths that the input data can traverse, determined by threshold tests at the bisections. Leaves represent the output values stored at the end points of branching. In RFR, a particular tree is grown in accordance with the realization of a random vector in order to introduce variation. The final prediction is based on aggregation over the ensemble of trees, referred to as the forest (Segal, 2004). On each of the trees, branches or nodes are made which are based on comparing a randomly selected feature to a random threshold. The randomness introduced in both variable selection and threshold determination has been shown to result in attractive properties such as a controlled variance, resistance to overtraining, and robustness to outliers as well as irrelevant variables. Moreover, RFR inherently provides estimates of generalization error and measures of variable importance (Bylander, 2002; Siroky, 2009). The process of dividing the input training data over branches are repeated until one or a pre-set number of data points are contained in each branch. This final node of the tree is referred to as a leaf, and it represents the outcome of that particular regression in the whole model. The structure of the forest and hence the RFR behaviour can be controlled by three parameters: the number of trees, the number of variables considered in each node (set to m=P/3, where P is the total number of predictor variables), and the number of data points that can reside in a leaf (our default value is 10). Having a very high number of leaves in the model can cause overfitting, which can be overcome by pruning, i.e. limiting the number of data points in each leaf. Increasing the number of trees in the forest has two main effects: The computation load will increase. An initial increase in the

accuracy of the regression will also be observed, before reaching a saturation point (Luppino et al., 2018), after which improvements are limited by a strong correlation between the trees (Breiman, 2001).

2.6 Pre-processing and input data for the model

The regression algorithm presented in Section 2.5 requires a training dataset for training the model and a test dataset to validate the trained model. In this study, the main inputs to the model are the surface solar radiations from CLARA and ERA5. In addition to these, clear sky indices were obtained by using shortwave surface radiation downward clear-sky (SWSDC) from ERA5 and GHI from ground measurements. By using clear sky indices, the RFR algorithm can take advantage of the sky stratification in different conditions. The daytime averages of solar zenith angle were also used as an input as it can provide the regression algorithm with the variation in solar elevation and its effects on surface radiations. Furthermore, latitudes and altitudes of the locations were used as input to the algorithm. In the training phase, 20% of randomly selected data was used from Norwegian locations, while the rest of the 80% data and data from Swedish locations were used in testing phase for validation of the model. The size of the training data was selected after running multiple runs with different sizes of data. Using more than 20% of data did not result in significant improvements. The model was tested with a number of trees ranging from 32 to 256 and pruning from 1 to 10 data points per leaf node. After multiple runs, 128 trees were selected with 10 data points per leaf node. The results presented in the next section are for the whole dataset.

Two main pre-processing procedures were applied in the training data of the regression model. Because of problems with convergence of the regression model, the missing data in CLARA and ground measurements was treated. First, training data with missing values in the ground measurements were discarded. This step eliminates the missing values in the ground data so that the regression model can converge, and also reduces the number of missing values in CLARA. This process was not performed on the test dataset, as missing values in the ground-measured data used in validation would not affect the errors statistics. Following previous studies that have shown that reanalyses can be used to fill the gaps in satellite datasets, we replaced in the second step the missing values of CLARA by corresponding values from ERA5 (Babar et al., 2018b; Urraca et al., 2017b; Urraca et al., 2018). These pre-processing steps enable the regression model to converge although with less training data.

2.7 Validation

In order to evaluate the performance of the RFR model, we introduce some common statistical measures. We first introduce the deviation (sometimes called error or residual) as the difference between the estimated (or predicted) and the observed global horizontal irradiance: $\delta = GHI_{estimated,i} - GHI_{observed,i}$, where the subscript i is a data point index.

A widely used measure of dispersion is the root mean square deviation (RMSD), computed from a sample of N data points as

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(GHI_{estimated,i} - GHI_{observed,i})^2}\,. \quad (1)$$

This measure combines both accuracy and precision.

The bias (or mean deviation) is used in the evaluation to quantify under- or overestimation. The bias is a measure of accuracy and is computed from the sample as

$$Bias = \frac{1}{N}\sum_{i=1}^{N}\left(GHI_{estimated,i} - GHI_{observed,i}\right) = \overline{GHI_{estimated}} - \overline{GHI_{observed}}\ , \quad (2)$$

where $\overline{GHI_{estimated}}$ and $\overline{GHI_{observed}}$ are the sample means of the estimated and the observed GHI values, respectively.

The mean absolute deviation (MAD) is another measure of dispersion, which give less weight to and is therefore less sensitive to outliers than the RMSD (and the variance). The sample MAD is computed as (Sanchez-Lorenzo et al., 2013; Willmott and Matsuura, 2005)

$$MAD = \frac{1}{N}\sum_{i=1}^{N}\left|GHI_{estimated,i} - GHI_{observed,i}\right|. \quad (3)$$

Following Karlsson et al. (2017), the standard deviation of δ (STD) is also used in the evaluation. The sample STD is computed as

$$STD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(\left(GHI_{estimated,i} - GHI_{observed,i}\right) - \left(\overline{GHI_{estimated}} - \overline{GHI_{observed}}\right)\right)^2}. \quad (4)$$

In addition, a bias-variance decomposition was used to obtain the optimal configuration of the RFR, with respect to the number of trees and the number of leaves. Moreover, R^2 and scatter plots are used to indicate the spread and overall correlation of the datasets with ground measurements.

3. Results

Table 1 compares performance of the models in terms of RMSD, MAD and bias for CLARA, ERA5 and the proposed RFR model. The RFR model performs better than the models that were used to construct it.

We start by looking at accuracy. For monthly averages of GHI at Norwegian locations, CLARA and ERA5 produced a bias of -1.6 Wm^{-2} and 3.9 Wm^{-2}, respectively. The RFR model delivered a bias of $-0.2\ Wm^{-2}$. The underestimation of the satellite model and the overestimation of the reanalysis is in agreement with previous studies (Babar et al., 2018a; Babar et al., 2018b; Urraca et al., 2017b; Urraca et al., 2018). The regression model underestimates the GHI, but the magnitude of the bias is reduced with 88% with respect to CLARA and with 95% with respect to ERA5, proving that the RFR model substantially improves the accuracy. These percentages are, as we will see, somewhat exaggerated when compared to seasonal values of the bias. Nonetheless, the seasonal biases are also much improved. The underestimation of the RFR model indicates that it weights CLARA higher than ERA5 on the whole, although the algorithm clearly adapts to exploit the strengths of either source under different conditions, as we will discuss below.

Regarding the dispersion measures, CLARA and ERA5 gave an MAD of 6.3 Wm^{-2} and 7.0 Wm^{-2}, respectively. The RFR model produced an MAD of 4.3 Wm^{-2}, which is a relative improvement of 32% and 39% with respect to CLARA and ERA5. Similarly, an RMSD of 6.6 Wm^{-2} was observed for the RFR model, while the RMSD of CLARA and ERA were 9.6 Wm^{-2} and 10.2 Wm^{-2}, respectively. The relative improvement in the RMSD was 31% and 35%, respectively. From the bias-variance decomposition of mean squared error ($MSE = RMSD^2$), the variance can be computed as: $Var = RMSD^2 - Bias^2$. We can use this to use that the variances of CLARA and ERA5 are very similar, and the variance of the RFR model is less half of these. This proves that the RFR model also provides a large improvement in precision. Table 1 also lists bias, MAD and RMSD for daily averages of GHI that show similar patterns as for the monthly averages.

Table 1 lists the error metrics after geographically grouping the ground measurement sites as explained in section 2.3. A brief overview of Table 1 shows that the proposed regression model improved all the four groups (above 65°N, below 65°N, coastal and inland). Like CLARA and ERA5, the proposed RFR

model performed better at above 65°N than below 65°N. Nevertheless, the accuracy and precision is improved in both of these groups.

Table 1: The RMSD, MAD and bias of the input datasets and the presented model are shown. The error metrics for all locations in addition to providing an analysis on below 65°N, above 65°N, coastal and inland locations are shown. Numbers without parentheses are monthly averaged errors while those in parentheses are daily averaged errors. Best results are indicated in bold.

	RMSD (Wm^{-2})			MAD (Wm^{-2})			Bias (Wm^{-2})		
	CLARA	ERA5	RFR	CLARA	ERA5	RFR	CLARA	ERA5	RFR
NIBIO sites	9.6 (19.1)	10.2 (26.7)	**6.6 (15.7)**	6.3 (13.1)	7.0 (16.7)	**4.3 (10.2)**	-1.6 (-2.0)	3.9 (3.9)	**-0.2 (-0.2)**
Above 65ºN	9.6 (16.0)	10.1 (26.3)	**6.5 (13.7)**	6.3 (9.7)	6.9 (14.5)	**4.2 (8.2)**	-1.6 (-2.9)	3.8 (5.6)	**-0.2 (-0.1)**
Below 65ºN	9.7 (19.5)	12.7 (26.8)	**8.0 (15.9)**	6.5 (13.6)	9.4 (17.3)	**5.4 (10.5)**	-1.8 (-1.8)	5.7 (3.9)	**0.1 (-0.1)**
Coastal	9.7 (16.7)	10.1 (26.7)	**6.6 (14.8)**	6.4 (11.4)	7.0 (16.3)	**4.3 (9.4)**	-1.7 (-1.1)	3.8 (4.9)	**-0.2 (0.4)**
Inland	8.2 (20.8)	11.2 (26.7)	**6.6 (16.4)**	5.7 (14.4)	7.9 (17.5)	**4.6 (10.8)**	-0.6 (-2.6)	4.5 (3.4)	**0.1 (-0.4)**

In addition, a seasonal error analysis was performed after dividing the yearly time series in groups of three months, i.e. February to April in FMA, May to July in MJJ, August to October in ASO, and November to January in NDJ. This type of grouping was preferred in this analysis because most locations analysed in this study are high latitude locations and at such locations the spread of solar radiation density is not as uniform as at other regions closer to the equator. At high latitude locations, most of the sun hours occur in summer months and least sun hours occur in winter months. By having such a grouping, summer and winter seasons are analysed separately. The seasonal analyses in Table 2 shows that errors decreased in all of the seasonal groups with the RFR model. However, the largest improvements were seen in FMA and MJJ. An analysis of the results of CLARA and ERA5 in NDJ and FMA shows that ERA5 performed better than CLARA in this period. This is mainly because of the low solar elevation in winter months, which increases errors in satellite-based estimates. However, CLARA performed better than ERA5 in MJJ and ASO.

The RFR model improves the accuracy and precision through all seasons. Nonetheless, the seasonal analysis reveals some interesting features: The bias of the RFR model varies over the year. The model underestimates in winter and overestimates in summer. However, we see that the biases of CLARA and ERA5 also fluctuate, and the RFR model succeeds in maintaining a much lower bias throughout the year. We may take this as a sign that the RFR model is flexible and adaptive, and manages to weight the input datasets in an appropriate way and combine their strengths to obtain good performance under various conditions. When it comes to the dispersion measures, the values of the RFR model follow the pattern of CLARA and ERA and largely decrease over the year. The largest relative improvements are seen in the FMA quarter, when the RFR model produces a 25% improvement in RMSD and a 39% improvement in MDA with respect to ERA5 (the best alternative). The magnitude of the bias reduction also over 70% for both models. The seasonal improvements are lower than the improvement in monthly averaged values, but the RFR model has much more consistent performance over the year than the input datasets. This is evident if one studies and compares the ranges or totals of the seasonal error metrics for the three models.

Table 2: The seasonal error analysis of CLARA, ERA5 and the RFR model are shown here. Major improvements occur in the FMA and MJJ quarters. Numbers without parentheses are monthly averaged errors while those in parentheses are daily averaged errors. Best results are indicated in bold.

	RMSD (Wm^{-2})				MAD (Wm^{-2})				Bias (Wm^{-2})			
	FMA	MJJ	ASO	NDJ	FMA	MJJ	ASO	NDJ	FMA	MJJ	ASO	NDJ
CLARA	15.3 (21.4)	8.8 (21.9)	8.9 (15.8)	6.9 (11.0)	10.4 (14.7)	6.7 (16.5)	4.9 (11.0)	4.4 (5.3)	-6.9 (-8.3)	1.3 (1.2)	**1.4 (1.1)**	0.3 (-2.3)
ERA	12.9 (23.5)	14.3 (40.7)	9.8 (23.7)	6.2 (9.3)	9.2 (16.4)	11.4 (30.9)	6.7 (16.5)	2.7 (4.2)	7.0 (7.0)	7.2 (7.1)	2.0 (2.1)	0.2 (0.3)
RFR Model	**9.7 (15.9)**	**7.4 (21.2)**	**7.8 (14.4)**	**5.6 (8.8)**	**5.6 (11.1)**	**5.6 (15.9)**	**4.4 (10.0)**	**2.3 (3.9)**	**-1.8 (-1.7)**	**-0.1 (-0.2)**	1.5 (1.5)	**0.0 (0.0)**

Finally, the R^2 values and the standard deviation (STD) of the Norwegian locations is analysed. Values of the coefficient of determination, R^2, are computed from the ground-measured and model data. The standard deviation is a measure of the spread of the prediction errors around their mean value. Table X shows the R^2 values and standard deviation for all Norwegian locations, in addition to below 65°N, above 65°N, coastal and inland regions. The standard deviation in Table 3 has units of Wm^{-2}, whereas R^2 has no units. For standard deviation, the smaller the value, the better the model estimates and for R^2, the larger the value, the better are the estimates.

Table 3: The R^2 and error standard deviation analysis of CLARA, ERA5 and the proposed RFR model for Norwegian locations is shown here. The RFR model improves the estimates in all types of geographical categories. The units of the standard deviation (STD) is Wm^{-2} and R^2 is unit-less. Best results are indicated in bold.

	NIBIO sites		Above 65ºN		Below 65ºN		Coastal		Inland	
	R^2	STD	R^2	STD	R^2	STD	R^2	STD	R^2	STD
CLARA	0.96	23.8	0.96	18.4	0.95	25.0	0.97	21.1	0.95	25.9
ERA	0.92	26.9	0.89	28.5	0.92	26.7	0.91	27.1	0.92	26.7
RFR model	**0.97**	**16.0**	**0.97**	**15.3**	**0.97**	**16.1**	**0.97**	**15.3**	**0.97**	**16.5**

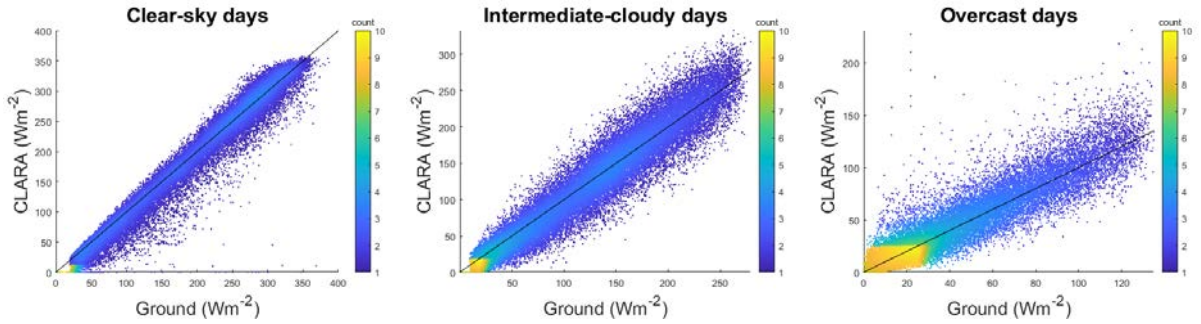
It can be observed that the proposed regression model improves the solar radiation estimates at all Norwegian locations. The largest improvements were observed in location above 65°N, although the differences are small. The proposed model had lower standard deviation than CLARA and ERA5 in all geographical groups. Note that CLARA performs better in coastal regions than in inland regions, while the opposite is true for ERA5.

3.1 Sky stratification in CLARA, ERA5 and the regression model

To evaluate the datasets for their performances in different sky conditions, the datasets were divided into clear-sky, intermediate-cloudiness and overcast categories. This division was established based on the clear-sky index (*Kc*), which is defined as the ratio of clear-sky GHI to the GHI recorded on the ground. Shortwave solar radiation clear-sky downwards (SWSCD) from ERA5 was used to obtain the clear-sky index. After calculating clear-sky index, *Kc*, following Smith et al. (2017) and Widén et al. (2017), values higher than 0.8 were considered as indicating a clear-sky day, values of *Kc* between 0.4 and 0.8 were considered as intermediate-cloudy, and values below 0.4 were considered as overcast. This kind of categorization is quite arbitrary, as days with *Kc* value of 0.8 or higher are not necessarily days with completely clear sky, but a majority of these days are expected to have a clear sky. This analysis is used here to roughly divide the sky conditions followed by a rigorous analysis. Any misclassification based on the clear sky indices will have similar effects on all the datasets.

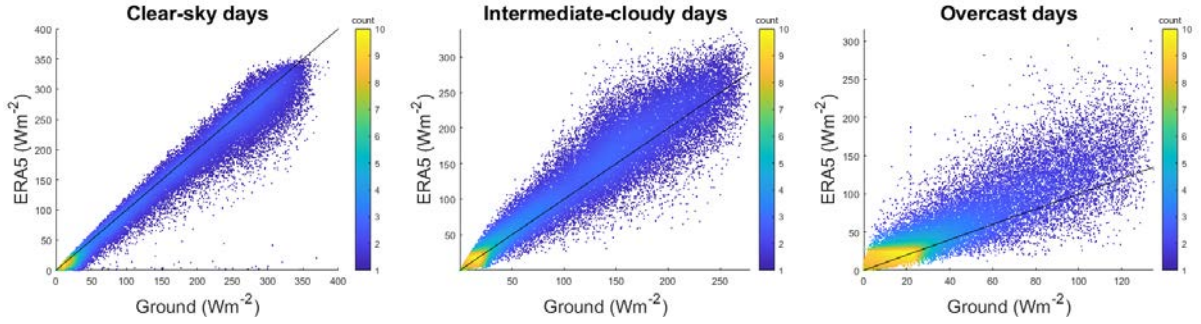
Figures 2-4 show the errors in the datasets under different sky categories. It can be seen from the figures and the tables that the RFR model improves the results in the clear-sky and intermediate cloudy

categories. However, in the overcast category, CLARA and the RFR model performed similarly besides that CLARA had a lower bias. On average, CLARA underestimated radiation in clear and cloudy conditions, while an overestimation was observed in overcast conditions. On the contrary, ERA5 overestimated radiation in cloudy and overcast conditions, while it was underestimated in clear-sky condition. ERA5 is reported to have a positive bias towards estimating days as clear sky and a negative bias towards estimating overcast days (Babar et al., 2018b). The reason for these biases is the higher concentration of total cloud water content in the ERA5 model on rather clear sky days and a lower concentration of total cloud water content in cloudy conditions. The underestimation in CLARA in clear sky and intermediate-cloudy days is possibly due to the use of an optically thick aerosol climatology – in this case the Global Aerosol Data Set/Optical Properties of Aerosols and Cloud (GADS/OPAC) climatology (Babar et al., 2018b; Mueller and Träger-Chatterjee, 2014). The RFR model underestimated solar radiation in clear sky condition and overestimated radiation in intermediate-cloudy and overcast conditions. Nevertheless, large improvements were observed in clear-sky and cloudy conditions. However, from a solar energy harvesting point of view, in overcast conditions smaller amounts of energy is produced as compared to clear-sky and intermediate-cloudy days.



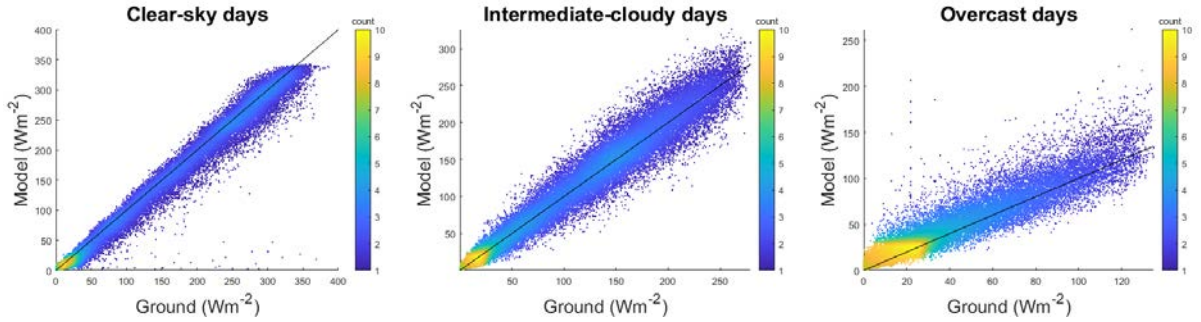
CLARA	RMSD (Wm^{-2})	MAD (Wm^{-2})	Bias (Wm^{-2})
Clear-sky	21.3	14.4	-7.1
Intermediate-cloudy	20.0	14.9	-2.8
Overcast	12.4	8.2	0.7

Figure 2: CLARA errors under clear-sky, intermediate-cloudy and overcast conditions for Norwegian sites. The scatter plots for different sky categories are also shown. The coloured legend bar shows the density of points.



ERA5	RMSD (Wm^{-2})	MAD (Wm^{-2})	Bias (Wm^{-2})
Clear-sky	25.0	15.9	-11.2
Intermediate-cloudy	28.2	19.4	9.5
Overcast	28.3	17.2	14.4

Figure 3: Same as Figure 2, but for ERA5.



RFR model	RMSD (Wm^{-2})	MAD (Wm^{-2})	Bias (Wm^{-2})
Clear-sky	17.4	11.3	-6.6
Intermediate-cloudy	16.8	11.8	1.7
Overcast	12.8	8.2	5.3

Figure 4: Same as Figure 2, but for the RFR model.

3.2 Testing the regression model on Swedish locations

In this section, the regression model is tested on five Swedish locations. Data from these locations were not used in the training of the model, therefore this analysis tests the robustness of the regression model proposed in this study. Table A2 in Appendix A lists the information on the Swedish locations used in the analysis.

Table 3 lists the errors for CLARA, ERA5 and the RFR model for individual Swedish locations. The errors for all locations are summarized in the last row of the table. In this analysis, it was found that the RFR model improved the solar radiation estimates for Swedish locations as well. The monthly MAD for all Swedish locations for CLARA and ERA5 was found to be 6.3 Wm^{-2} and 5.6 Wm^{-2}, respectively. At these locations, the RFR model gave a MAD of 4.5 Wm^{-2}. Similarly, the daily averages were also improved in the RFR model. As previously observed for Norwegian locations, CLARA underestimated the solar radiation and ERA5 overestimated it for Swedish locations. The proposed RFR model underestimated the solar radiation as well, but the magnitude of the bias was smaller than for CLARA and ERA5. This analysis shows that the proposed model can at least be used for Swedish locations that may have a similar climate in terms of cloud, snow and sunlight conditions.

Table 3: The RMSD, MAD and Bias of the input datasets and the RFR model for Swedish locations is shown here. These locations were not used in the training of the regression model. Numbers without parentheses are monthly averaged errors while those in parentheses are daily averaged errors. Best results are indicated in bold.

	RMSD (Wm^{-2})			MAD (Wm^{-2})			Bias (Wm^{-2})		
	CLARA	ERA5	RFR	CLARA	ERA5	RFR	CLARA	ERA5	RFR
Kiruna	17.2 (26.6)	**7.6** (24.0)	11.0 **(18.7)**	10.1 (16.6)	**4.9** (14.4)	6.8 **(11.7)**	-7.0 (-8.2)	**-2.3** **(-2.5)**	-5.9 (-6.0)
Luleå	10.6 (24.4)	10.4 (25.1)	**5.6** **(17.5)**	6.9 (14.9)	6.6 (15.3)	**3.8** **(11.0)**	-4.4 (-4.2)	5.1 (4.9)	**-2.1** **(-2.1)**
Umeå	8.3 (16.4)	7.1 (23.0)	**5.5** **(13.5)**	6.1 (11.5)	4.4 (14.2)	**3.8** **(9.1)**	-3.2 (-3.5)	**2.0** **(2.1)**	-2.6 (-2.5)
Stockholm	6.8 (16.4)	7.0 (23.6)	**5.9** **(14.6)**	5.1 (11.5)	4.8 (15.7)	**4.5** **(10.0)**	**2.6** **(2.5)**	3.1 (3.1)	3.9 (4.0)
Göteborg	**4.7** (14.9)	9.5 (26.1)	4.8 **(14.4)**	3.5 (10.5)	7.3 (17.0)	3.7 (9.9)	**1.6** **(1.8)**	6.9 (6.8)	3.0 (2.9)
SMHI locations	10.4 (20.3)	8.4 (24.4)	**6.9** **(15.9)**	6.3 (13.0)	5.6 (15.3)	**4.5** **(10.3)**	-2.1 (-2.3)	2.9 (2.9)	**-0.8** **(-0.7)**

4 Conclusion

Studies have shown that satellite estimation of solar radiation provide reasonable estimates and reanalyses can be used to fill the gaps when satellite datasets are not available or they contain missing data. It has also been observed that at high latitude locations there are a larger number of missing values in satellite-derived data, as in CLARA. Some previous studies have reported that prediction errors increase with latitude, so the available datasets have a systematic bias that grows with latitude. This study proposes a novel method to construct an improved dataset by combining a surface solar radiation dataset based on satellite measurements (CLARA-A2) and a newly published global reanalysis dataset (ERA5). The assumption used in this study is that the underestimation in satellite models and the overestimation in reanalyses can be largely cancelled and overcome if they are fused in a regression model to improve the estimates of surface solar radiation. The proposed regression model is constructed by using the random forest regression method, which is a machine learning algorithm based on regression trees and ensemble learning.

It is seen that on monthly and daily averages of radiation, the regression model provided more accurate estimations than CLARA and ERA5. On monthly averages of surface solar radiation for Norwegian locations, CLARA provided an MAD of 6.3 Wm^{-2} while ERA5 provided an MAD of 7.0 Wm^{-2}. The regression model reduced the error to a MAD of 4.3 Wm^{-2}. Similarly, on daily averages, CLARA and ERA5 provided MADs of 13.1 Wm^{-2} and 16.7 Wm^{-2}, respectively, while the regression model gave a MAD of 10.2 Wm^{-2}. Similar improvements were seen in RMSE values, proving that the RFR model has significantly improved precision with respect to the input datasets. In addition the RFR model was seen to provide large reductions in both annual and seasonal bias, showing that the accuracy improves as much as the precision.

A discussion of the seasonal analysis concluded that the RFR model succeeds in combining the input datasets in an adaptive fashion, such that the strengths of both models are exploited to produce consistently high performance under all conditions and throughout the whole year. Moreover, from a geographical analysis of errors it was observed that large improvements were obtained in locations above 65°N and coastal regions. A seasonal error analysis is performed and it is observed that the regression model provided better estimates than CLARA and ERA5 in all seasons of the year with large improvements in the period of November to April. A sky stratification analysis was performed on Norwegian locations to assess the datasets in different sky conditions. It was observed that the regression model improved solar radiation estimates in all sky condition, especially in clear-sky and intermediate-cloudy conditions. Additionally, in terms of standard deviation, large improvements were found inland and below 65°N. The proposed model was also tested on Swedish locations, that were not included in the training set, and very similar improvements were observed.

Overall, the regression model provides an improved alternative to the available reanalyses and satellite based estimates of surface solar radiation. In addition to an improved dataset, this study also highlights the important role of machine learning algorithms in the production of sophisticated databases for high latitude locations.

Appendix A

Table A1

Lists of Norwegian locations with their coordinates, altitudes and land type.

	Station	Latitude	Longitude	Altitude	Land type
1	Holt	69.65	18.91	12	Coastal
2	Sortland	68.65	15.28	14	Coastal
3	Vågønes	67.28	14.45	26	Coastal
4	Tjøtta	65.83	12.43	10	Coastal
5	Skogmo	64.51	12.02	32	Inland
6	Rissa	63.59	9.97	23	Coastal
7	Kvithamar	63.49	10.88	28	Inland
8	Skjetlein	63.34	10.3	44	Coastal
9	Surnadal	62.98	8.69	5	Inland
10	Tingvoll	62.91	8.19	23	Coastal
11	Fåvang	61.46	10.19	184	Inland
12	Fureneset	61.29	5.04	12	Coastal
13	Gausdal	61.22	10.26	375	Inland
14	Løken	61.12	9.06	527	Inland
15	Ilseng	60.8	11.2	182	Inland
16	Kise	60.77	10.81	129	Inland
17	Apelsvoll	60.7	10.87	262	Inland
18	Hønefoss	60.14	10.27	126	Inland
19	Årnes	60.13	11.39	162	Inland
20	Etne	59.66	5.95	8	Inland
21	Ås	59.66	10.78	94	Inland
22	Bø	59.42	9.03	105	Inland
23	Rakkestad	59.39	11.39	102	Inland
24	Ramnes	59.38	10.24	39	Coastal
25	Tomb	59.32	10.81	12	Coastal
26	Gjerpen	59.23	9.58	41	Coastal
27	Hjelmeland	59.23	6.15	43	Inland
28	Tjølling	59.05	10.13	19	Coastal
29	Særheim	58.76	5.65	90	Coastal
30	Landvik	58.34	8.52	10	Coastal
31	Lyngdal	58.13	7.05	4	Inland

Table A2

Lists of Swedish locations with their coordinates, altitudes and land type.

	Station	Latitude	Longitude	Altitude	Land type
1	Kiruna	67.83	20.43	408	Inland
2	Luleå	65.55	22.13	17	Coastal
3	Umeå	63.82	20.25	10	Coastal
4	Stockholm	59.35	18.07	30	Coastal
5	Goteborg	57.70	12.00	5	Coastal

Appendix B

Table B1

The following years are not included in the study.

	Station	Years having more than 5% missing data	Years failing Long and Dutton test	Years having operational error (snow/frost/ shading/soiling)	Years having equipment error
1	Holt	2001,2002,2006,2007,2008,2010	2013		2000
2	Sortland	2000,2006,2007,2010,2013			
3	Vågønes	2006,2007		2002	
4	Tjøtta	2006,2007			2008, 2012
5	Skogmo	2006,2007,2008,2015		2011	2013, 2014
6	Rissa	2006,2007	2000		
7	Kvithamar	2006,2007,2013			
8	Skjetlein	2006,2007	2000		
9	Surnadal	2006,2007,2014			
10	Tingvoll	2006,2007,2012			
11	Fåvang	2006,2007			2001
12	Fureneset	2006,2007,2011,2012			
13	Gausdal	2006,2007,2009			2015
14	Løken	2006,2007			
15	Ilseng	2006,2007,2004	2000	2009	
16	Kise	2002,2006,2007,2015		2013	
17	Apelsvoll	2006,2007		2002,2003,2004	2009
18	Hønefoss	2006,2007	2000		
19	Årnes	2006,2007			
20	Etne	2006,2007		2004,2012	
21	Ås	2006,2007			
22	Bø	2000,2006,2007			
23	Rakkestad	2006,2007			
24	Ramnes	2006,2007		2009	
25	Tomb	2006,2007	2009		
26	Gjerpen	2006,2007,2015			
27	Hjelmeland	2006,2007			2002, 2015
28	Tjølling	2006,2007,2008,2014		2012,2015	2009, 2010
29	Særheim	2000,2006,2007			
30	Landvik	2006,2007		2005,2010,2014, 2015	
31	Lyngdal	2006,2007	2001		

References

Anderson, D.W., Kish, L., Cornell, R.G., 1980. On stratification, grouping and matching. J Scandinavian Journal of Statistics, 61-66.

Babar, B., Graversen, R., Boström, T., 2018a. Evaluating CM-SAF solar radiation CLARA-A1 and CLARA-A2 datasets in Scandinavia. Solar Energy 170, 76-85.

Babar, B., Rune, G., Tobias, B., 2018b. Solar radiation estimation at high latitudes: Assessment of the CMSAF databases, ASR and ERA5. under-review in Solar Energy.

Bagirov, A.M., Mahmood, A., Barton, A., 2017. Prediction of monthly rainfall in Victoria, Australia: Clusterwise linear regression approach. J Atmospheric Research

188, 20-29.

Breiman, L.J.M.l., 2001. Random forests. Machine learning 45(1), 5-32.

Bylander, T.J.M.L., 2002. Estimating generalization error on two-class datasets using out-of-bag estimates. Machine Learning 48(1-3), 287-297.

Cano, D., Monget, J.-M., Albuisson, M., Guillard, H., Regas, N., Wald, L., 1986. A method for the determination of the global solar radiation from meteorological satellites data. Solar energy 37(1), 31-39.

Cebecauer, T., Šúri, M., 2012. Correction of satellite-derived DNI time series using locally-resolved aerosol Data, Proceedings of the SolarPACES Conference, Marrakech, Morocco.

Davy, R.J., Huang, J.R., Troccoli, A.J.S.E., 2016. Improving the accuracy of hourly satellite-derived solar irradiance by combining with dynamically downscaled estimates using generalised additive models. Solar Energy 135, 854-863.

Dee, D.P., Uppala, S.M., Simmons, A., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., Bauer, d.P., 2011. The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. Quarterly Journal of the royal meteorological society 137(656), 553-597.

ECMWF, 2018. ERA5 data documentation.

Gueymard, C.A.J.S.E., 2012. Temporal variability in direct and global irradiance at various time scales as affected by aerosols. Solar Energy 86(12), 3544-3553.

Hersbach, H., Dee, D., 2016. ERA5 reanalysis is in production. ECMWF newsletter 147(7).

Hsu, D.J.A.e., 2015. Comparison of integrated clustering methods for accurate and stable prediction of building energy consumption data. Applied energy 160, 153-163.

Karlsson, K.-G., Anttila, K., Trentmann, J., Stengel, M., Meirink, J.F., Devasthale, A., Hanschmann, T., Kothe, S., Jääskeläinen, E., Sedlar, J., 2017. CLARA-A2: the second edition of the CM SAF cloud and radiation data record from 34 years of global AVHRR data. Atmospheric Chemistry and Physics 17(9), 5809.

Kennedy, A.D., Dong, X., Xi, B., Xie, S., Zhang, Y., Chen, J., 2011. A comparison of MERRA and NARR reanalyses with the DOE ARM SGP data. Journal of Climate 24(17), 4541-4557.

Long, C.N., Dutton, E.G., 2010. BSRN Global Network recommended QC tests, V2. x.

Luppino, L.T., Bianchi, F.M., Moser, G., Anfinsen, S.N., 2018. Remote sensing image regression for heterogeneous change detection. arXiv preprint arXiv:1807.11766.

Mueller, R., Matsoukas, C., Gratzki, A., Behr, H., Hollmann, R.J.R.S.o.E., 2009. The CM-SAF operational scheme for the satellite based retrieval of solar surface irradiance—A LUT based eigenvector hybrid approach. Remote Sensing of Environment 113(5), 1012-1024.

Mueller, R., Träger-Chatterjee, C., 2014. Brief accuracy assessment of aerosol climatologies for the retrieval of solar surface radiation. Atmosphere 5(4), 959-972.

Müller, R., Pfeifroth, U., Träger-Chatterjee, C., Trentmann, J., Cremer, R., 2015. Digging the METEOSAT treasure—3 decades of solar surface radiation. Remote Sensing 7(6), 8067-8101.

NIBIO, 2018.

Noia, M., Ratto, C., Festa, R., 1993. Solar irradiance estimation from geostationary satellite data: II. Physical models. Solar Energy 51(6), 457-465.

Persson, T., 2000. Measurements of solar radiation in Sweden 1983-1998. na.

Pfeifroth, U., Kothe, S., Müller, R., Trentmann, J., Hollmann, R., Fuchs, P., Werscheck, M., 2017. Surface Radiation Data Set–Heliosat (SARAH)–Edition 2, Satellite Application Facility on Climate Monitoring.
Pinker, R., Laszlo, I., 1992. Modeling surface solar irradiance for satellite applications on a global scale. Journal of Applied Meteorology 31(2), 194-211.
Polo, J., Martín, L., Vindel, J.J.R.E., 2015. Correcting satellite derived DNI with systematic and seasonal deviations: application to India. Renewable Energy 80, 238-243.
Polo, J., Wilbert, S., Ruiz-Arias, J.A., Meyer, R., Gueymard, C., Suri, M., Martin, L., Mieslinger, T., Blanc, P., Grant, I., 2016. Preliminary survey on site-adaptation techniques for satellite-derived and reanalysis solar radiation datasets. Solar Energy 132, 25-37.
Rigollier, C., Lefèvre, M., Wald, L., 2004. The method Heliosat-2 for deriving shortwave solar radiation from satellite images. Solar Energy 77(2), 159-169.
Riihelä, A., Carlund, T., Trentmann, J., Müller, R., Lindfors, A.V., 2015. Validation of CM SAF surface solar radiation datasets over Finland and Sweden. Remote Sensing 7(6), 6663-6682.
SAF, C., 2016. 2: Algorithm Theoretical Basis Document–CM SAF Cloud, Albedo, Radiation data record, AVHRR-based, Edition 2 (CLARA-A2)–Cloud Fraction. SAF/CM/DWD/ATBD/CMA_AVHRR version 2.0, https://doi. org/10.5676/EUM_SAF_CM ….
Sanchez-Lorenzo, A., Wild, M., Trentmann, J.J.R.S.o.E., 2013. Validation and stability assessment of the monthly mean CM SAF surface solar radiation dataset over Europe against a homogenized surface dataset (1983–2005). Remote Sensing of Environment 134, 355-366.
Schmetz, J., Pili, P., Tjemkes, S., Just, D., Kerkmann, J., Rota, S., Ratier, A., 2002. An introduction to Meteosat second generation (MSG). Bulletin of the American Meteorological Society 83(7), 977-992.
Schulz, J., Albert, P., Behr, H.-D., Caprion, D., Deneke, H., Dewitte, S., Durr, B., Fuchs, P., Gratzki, A., Hechler, P., 2009. Operational climate monitoring from space: the EUMETSAT Satellite Application Facility on Climate Monitoring (CM-SAF). Atmospheric Chemistry and Physics 9(5), 1687-1709.
Segal, M.R., 2004. Machine learning benchmarks and random forest regression.
Sengupta, M., Habte, A., Gueymard, C., Wilbert, S., Renne, D., 2017. Best practices handbook for the collection and use of solar resource data for solar energy applications. National Renewable Energy Lab.(NREL), Golden, CO (United States).
Siroky, D.S.J.S.S., 2009. Navigating random forests and related advances in algorithmic modeling. Statistics Surveys 3, 147-163.
Smith, C.J., Bright, J.M., Crook, R., 2017. Cloud cover effect of clear-sky index distributions and differences between human and automatic cloud observations. Solar Energy 144, 10-21.
Späth, H.J.C., 1979. Algorithm 39 Clusterwise linear regression. Computing 22(4), 367-373.
Stoffel, T., Renne, D., Myers, D., Wilcox, S., Sengupta, M., George, R., Turchi, C., 2010. Concentrating solar power. Golden: National Renewable Energy Laboratory.
Suri, M., Cebecauer, T., 2014. Satellite-based solar resource data: Model validation statistics versus user's uncertainty, ASES SOLAR 2014 Conference, San Francisco. pp. 7-9.
Tarpley, J., 1979. Estimating incident solar radiation at the surface from geostationary satellite data. Journal of Applied Meteorology 18(9), 1172-1181.
Tso, G.K., Yau, K.K.J.E., 2007. Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. Energy 32(9), 1761-1768.
Urraca, R., Gracia-Amillo, A.M., Huld, T., Martinez-de-Pison, F.J., Trentmann, J., Lindfors, A.V., Riihelä, A., Sanz-Garcia, A., 2017a. Quality control of global solar radiation data with satellite-based products. Solar Energy 158, 49-62.
Urraca, R., Gracia-Amillo, A.M., Koubli, E., Huld, T., Trentmann, J., Riihelä, A., Lindfors, A.V., Palmer, D., Gottschalg, R., Antonanzas-Torres, F., 2017b. Extensive validation of CM SAF surface radiation products over Europe. Remote sensing of environment 199, 171-186.
Urraca, R., Huld, T., Gracia-Amillo, A., Martinez-de-Pison, F.J., Kaspar, F., Sanz-Garcia, A., 2018. Evaluation of global horizontal irradiance estimates from ERA5 and COSMO-REA6 reanalyses using ground and satellite-based data. Solar Energy 164, 339-354.

Widén, J., Shepero, M., Munkhammar, J., 2017. On the properties of aggregate clear-sky index distributions and an improved model for spatially correlated instantaneous solar irradiance. Solar Energy 157, 566-580.
Wild, M., 2008. Short-wave and long-wave surface radiation budgets in GCMs: A review based on the IPCC-AR4/CMIP3 models. Tellus A: Dynamic Meteorology and Oceanography 60.5 60(5), 932-945.
Wild, M., Folini, D., Henschel, F., Fischer, N., Müller, B., 2015. Projections of long-term changes in solar radiation based on CMIP5 climate models and their influence on energy yields of photovoltaic systems. Solar Energy 116, 12-24.
Willmott, C.J., Matsuura, K.J.C.r., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate research 30(1), 79-82.
You, Q., Sanchez-Lorenzo, A., Wild, M., Folini, D., Fraedrich, K., Ren, G., Kang, S., 2013. Decadal variation of surface solar radiation in the Tibetan Plateau from observations, reanalysis and model simulations. Climate dynamics 40(7-8), 2073-2086.
Yu, Z., Haghighat, F., Fung, B.C., Yoshino, H.J.E., Buildings, 2010. A decision tree method for building energy demand modeling. 42(10), 1637-1646.
Zhao, L., Lee, X., Liu, S., 2013. Correcting surface solar radiation of two data assimilation systems against FLUXNET observations in North America. Journal of Geophysical Research: Atmospheres 118(17), 9552-9564.