Bachelor thesis

Global verification of incoming shortwave radiation for several weather forecast models and different model combinations

Alexandra Reiß
Student ID: 4735637

December 14th, 2021

Albert-Ludwigs-Universität Freiburg
Chair of Environmental Meteorology
In cooperation with meteoblue AG Basel

Supervisor

Prof. Dr. Dirk Schindler

Prof. Dr. Carsten Dormann

Dr. Sebastian Schlögl (External)

Abstract

This thesis deals with solar radiation verification of four global (NEMSGLOBAL, GFS, ICON and MFGLOBAL) weather forecast models and the reanalysis model ERA5. With focusing on incoming shortwave radiation, the performances of these models are evaluated and compared for the years 2018 until 2020. The goal of this thesis is to find out which models perform the best and how the forecast error can be lowered by implementing a multi-model approach. Furthermore, seasonal, and spatial patterns are investigated. To achieve that, a quality control procedure is applied to measured data based on 81 stations of chosen weather station networks. For the verification, measured data are compared to forecast data provided by meteoblue AG by calculating several statistical metrics. Among the raw models, ERA5 performed the best. Results of the multi-models show a significant reduction of the forecast error by up to 40 % when combining two or more models. Within the best performing multi-models, ICON was usually weighted the highest (up to 50 %). Several seasonal inconsistencies were observed, especially when four models were combined. Spatial analysis shows that ICON and GFS perform highly variable throughout the globe, while NEMSGLOBAL and MFGLOBAL perform rather consistent. These findings were as well represented by regional variabilities of the weightings of the models. Results indicate that the multi-models form a potential approach for improving solar irradiation forecast, however, showing seasonal and spatial inconsistencies. That shows the potential of further investigation of forecast models and their combination, especially on a regional basis.

Kurzzusammenfassung

Diese Arbeit befasst sich mit einer Überprüfung der Vorhersage von Solarstrahlung von vier globalen Wettervorhersagemodellen (NEMSGLOBAL, GFS, ICON und MFGLOBAL) und dem Reanalysemodell ERA5. Mit Fokus auf die einfallende kurzwellige Strahlung wird die Leistung dieser Modelle für die Jahre 2018 bis 2020 bewertet und verglichen. Ziel dieser Arbeit ist es, herauszufinden, welche Modelle am besten abschneiden und wie der Vorhersagefehler durch einen Multi-Modell-Ansatz gesenkt werden kann. Außerdem werden saisonale und räumliche Muster untersucht. Dafür wird ein Qualitätskontrollverfahren auf Messdaten von 81 Stationen ausgewählter Datennetzwerke angewendet. Zur Verifizierung werden die gemessenen Daten mit den Vorhersagedaten, bereitgestellt von meteoblue AG, mit Hilfe von verschiedenen statistischen Metriken verglichen. Unter den Rohmodellen schnitt ERA5 am besten ab. Die Ergebnisse der Multimodelle zeigen eine signifikante Reduktion des Vorhersagefehlers um bis zu 40 %, wenn zwei oder mehr Modelle kombiniert werden. Innerhalb der leistungsstärksten Multimodelle wurde ICON in der Regel am höchsten gewichtet (bis zu 50 %). Es wurden mehrere saisonale Unstimmigkeiten beobachtet, insbesondere wenn vier Modelle kombiniert wurden. Die räumliche Analyse zeigt, dass ICON und GFS auf der ganzen Welt sehr unterschiedlich abschneiden, während der Vorhersagefehler von NEMSGLOBAL und MFGLOBAL global gesehen weniger variiert. Diese Ergebnisse wurden auch durch die regionale Variabilität der Gewichtung der Modelle repräsentiert. Die Ergebnisse deuten darauf hin, dass die Multimodelle ein potenzieller Ansatz zur Verbesserung der Vorhersage der Solarstrahlung sind, während sie allerdings jahreszeitliche und räumliche Inkonsistenzen aufzeigen. Dies zeigt das Potenzial weiterer Untersuchungen von Vorhersagemodellen und ihrer Kombination, insbesondere auf regionaler Basis.

Acknowledgements

At this point I would like to give special thanks to Dr. Sebastian Schlögl as he continuously supported the process of this work and provided valuable input. Furthermore, I want to thank Michael Bührer for his patience and effort in finding solutions for problems that occurred during data analyses. A sincere thanks goes to all members of meteoblue AG, especially to the meteorological department, who have welcomed me into the team over the last few months and made this time an enriching and experiential one. Finally, I thank my family and friends for always keeping me motivated.

Contents

1 Introduction

Forecasting is useful for many situations. Because forecasting means to predict the future as accurately as possible, it is an important aid to effectively and efficiently plan and make decisions (Antonio et al., 2018). Among the public, weather forecasts are probably the most visible and commonly used type of scientific prediction. For a long time, much effort has been made to estimate the uncertainty associated with weather on a day-to-day basis (Roulston et al., 2006). Next to the importance of predicting severe weather conditions like storms, heavy rainfall, or tornados to prevent high destruction, weather forecasting can be a useful tool for agriculture. Agriculture is highly dependent on climate. As a result, crop yield variability is affected by year-to-year climatic variability. The usefulness for climatic knowledge is evident for many sectors (Cantelaube & Terres, 2005). The sector most commonly associated with forecasting is the energy sector. One of the major challenges for future global energy supply will be the integration of renewable energy sources (Heinemann et al., 2006a). The rapidly evolving situation in the energy market leads to the need for research on solar power predictions (Perez et al., 2013). Therefore, an increasing interest in precise and applicable modeling, forecasting and prediction of solar irradiance has evolved (Wang et al., 2012). To efficiently plan and operate solar energy systems, forecasts for up to 48 h have to be provided (Heinemann et al., 2006b). Numerical weather prediction (NWP) is one of the best tools for hour- and day-ahead forecasts (Kleissl, 2010). NWP's infer local cloud information and indirectly transmitted radiation, while the dynamic of the atmosphere is modeled through complex physical equations (Perez et al., 2013). For decision-makers, it is of crucial importance to examine and compare the potential and performance of NWP-models. (Huang & Thatcher, 2017). Within verification of models, measured data are compared with model forecast data. Measured data provide empirical evidence but are often subject to significant measurement uncertainty (Yang et al., 2018). To obtain reliable results of NWP's performance, measured data need to be analyzed and checked for quality (Perez-Astudillo et al., 2019). In literature, various methods for quantifying the performance of NWP's have been established, as well as approaches to improve the forecast certainty. For example, ensemble forecasting, in which multiple simulations of the atmosphere reflect the uncertainty of the models' outputs, has become a standard tool for operational weather forecasting (Roulston et al., 2006). Further approaches have been mentioned, in which the output of different models was averaged and combined, with the consequence of better performance than individual models (Perez et al., 2013). Results like this show the potential of using NWPs in new approaches to further improve the forecast of solar irradiation.

2 Research question

With climate change and the growing need for renewable energies, efficient photovoltaic systems and accurate forecast of available solar irradiance are of major importance. Even though growing research to improve forecasts can be observed, the availability of ground measurements of certain regions is still insufficient. Numerical weather forecasts form a widely used alternative by predicting global radiation data. Different models exist on the market, each with their own global and regional strengths and weaknesses of predicting processes in the atmosphere. To support decision-makers, the estimation of the performance of different weather models is mandatory. Measurements used for verification are accompanied by uncertainties due to several error sources. Unfortunately, no general recommended quality control procedure exists, which makes a comparison of different research difficult. To improve forecasts, new approaches such as multi-modeling, in which the output of several individual models is combined, have been suggested. Yet, they have been insufficiently investigated. With applying well-known quality control procedures to measured solar irradiation data and conducting raw model verifications, this thesis gives answers to the following questions: How many percent of all measured data do not pass the quality control? Which raw models perform on average the best, and how did the model accuracy change within the last years? By giving an insight of multi-models and investigating spatial and seasonal differences, answers are given to: Does a multi-model mix reduce the solar radiation forecast error? Which multi-model performs best and how much could the error be lowered? Is the model performance sensitive to different locations or different time periods within the year?

3 State of the art

3.1 Numerical weather forecast

The most accurate forecasting of solar radiation has been generated through NWPs in the past (Huang & Thatcher, 2017; Mathiesen & Kleissl, 2011). It becomes the best approach beyond several hours in advance, while it systematically simulates key atmospheric processes and their evolution at large scales (Huang et al., 2018). NWP's process prevailing weather observations assimilated into the model's framework. They are used to produce predictions for a variety of meteorological elements, such as radiation (NCEI, 2021). With highly complicated equations, essential physical processes in the atmosphere are calculated on powerful computers. Numerical methods are used to calculate the temporal evolution of the model's variables in a three-dimensional spatial grid extending from the ground to a top boundary. The horizontal distance between two neighboring grid points (mesh size) is an important parameter. The smaller it is, the more detailed the forecast model can predict the conditions of the earth's surface and atmosphere. In addition, the vertical layer thickness can vary from a few meters near the ground to several hundred meters. Different models are distinguished by these parameters, whereas the accuracy highly varies, especially for different regions (DWD, 2021c). The performance of NWP depends on individual models and locations (Huang et al., 2018). Forecast errors often arise from an incorrect representation of convective clouds, because NWP-models do not precisely predict the stochastic nature of clouds at a small spatial and temporal scale (Huang & Thatcher, 2017). Therefrom, NWP-models usually over- or underpredict several climatologic parameters. Moreover, model uncertainties result from approximation-errors of physical parameterizations (Boisserie et al., 2014), as well as of determining the initial conditions the model is based on (Troccoli, 2010). Many popular NWP's have been significantly improved in conventional operational performance due to achievements in research on the one hand, and advancements in computation capability on the other hand (Huang et al., 2018). In the need for improving forecasting approaches, literature gives first insights into the advantages of combining models. In the study of Perez et al (2013), after evaluating raw model performances, the best performing raw models were blended. Results showed a slightly better performance compared to the results of the individual models. This approach was investigated by Huang et al. (2018) for several sites in Australia, statistically confirming, that forecast is improved by blending multiple models. While the randomness in the forecast irradiances can be statistically averaged, different models can partly compensate for each other (Gregory et al., 2012; Perez et al., 2013). These results give first insights into the

potential of multi-models to improve solar irradiance forecasting, particularly on a regional basis. To shed light on the variety of different NWP-models, several well-known models will be described shortly in the following.

3.1.1 Global Forecast System (GFS)

GFS is a global forecasting model established by the National Oceanic and Atmospheric Administration (NOAA) through National Centers for Environmental Prediction (NCEP). Its forecast is published at 00, 06, 12 and 18 of Coordinated Universal Time (UTC). GFS considers certain parameters like absorption effects from water vapor, ozone, oxygen, and methane as well as cloud optical depth, albedo and more (Mathiesen & Kleissl, 2011). Its horizontal resolution is approximately 13 km, and the vertical is divided by 127 layers (NCEP, 2021).

3.1.2 NOAA Environment Monitoring System Global (NEMS Global)

The NEMSGLOBAL is the global component of the multi-scale model family NEMS that significantly improves cloud development and precipitation forecast. It was published by meteoblue AG and is the successor of Nonhydrostatic Mesoscale Models (NMM). Its horizontal resolution is 30 km (meteoblue AG, 2021b).

3.1.3 Icosahedral Nonhydrostatic Model (ICON)

ICON is a numerical weather prediction model designed by the German Weather Service (DWD) in cooperation with the Max Planck Institute for Meteorology. It was the first model using an icosahedral grid. The global ICON grid has 2,949,120 triangles, corresponding to an average area of 173 km² and thus to an effective mesh size of about 13 km (DWD, 2021a). It has 90 vertical layers. For Europe, ICON owns one refined subregion ("nest"), which leads to a higher regional resolution (Reinert et al., 2021).

3.1.4 Meteofrance Global (MFGLOBAL)

MFGLOBAL, also called ARPEGE40, is a weather model by the French national weather service. The global resolution is 40 km. It has 105 vertical levels. Daily forecasts are made at 00, 06, 12 and 18 UTC (CNRM, 2021).

3.1.5 ECMWF Reanalysis 5th Generation (ERA5)

The ERA5 model is, in contrast to the other models, a global reanalysis model. It was produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) (Hersbach et al., 2020). ECMWF periodically uses this forecast model and data assimilation systems to "reanalyze" archived observations, through which global datasets are created (ECMWF, 2021a). It provides a great number of outputs that are available in an hourly resolution. (Hersbach et al., 2020) It consists of 137 levels in the vertical and has a spatial resolution of 31 km (ECMWF, 2021b).

3.2 Solar irradiance and measurements

For weather analyses, forecasts, and weather warnings as well as research, meteorological observations, and environmental measurements are necessary. The need to provide high-quality meteorological data has led to a great evolution of automatic data acquisition systems, in which data are saved in large databases. Therefore, over the last 20 years, the number of automated weather station networks has greatly increased (Estévez et al., 2011). The Basic Surface Radiation Network (BSRN) is the central archive under the World Radiation Monitoring Center (WRMC) (Yang et al., 2018). It has been established to provide high-quality radiation measurements, aiming at detecting important changes in the surface balance. Therefore, is has been widely used in scientific applications (Roesch et al., 2011). Another well-known central depository for solar radiation data is the World Radiation Data Center (WRDC) located in Russia. Member countries of the World Meteorological Organization (WMO) contribute measured data of over 1000 globally distributed measurement sites (Badescu, 2008). Measuring solar radiation by ground-based weather stations is one of the most accurate methods (Alani et al., 2021). Direct radiation (*DIR*), also called beam irradiance, is the radiation coming directly from the sun. It is usually measured with a pyrheliometer. Diffuse radiation (*DIF*) is the sunlight scattered in the atmosphere, measured with a shaded pyranometer, therefore, being shielded from the *DIR*. Together, these fractions sum up to the global horizontal irradiance (*GHI*). The latter can also be directly measured with an unshaded pyranometer.

Measured data are always accompanied by significant measurement uncertainty (Yang et al., 2018). Especially measurements of solar radiation are prone to errors (Journée & Bertrand, 2011). These errors arise not only through inaccuracies and imprecision of an instrument, but also through insufficient data management (Behar et al., 2015). Next to power outages or communication problems causing

storage problems or data gaps (Alani et al., 2021), the most common error sources are sensors and their construction. These include, for example, the cosine response (Muneer et al., 2007). Ideally, a pyranometer has a directional response to the cosine law. Lambert's cosine law states that the luminous flux produced by a focused beam on a flat surface is proportional to the cosine of the angle of incidence of the beam on the surface (Badescu, 2008). However, this response is influenced by several factors such as the detector or the construction of the domes. Due to the high sensitivities of the sensor, the error increases, the lower the sun angle (Alani et al., 2021). That is why erroneous measurements can be observed especially during sunrise and sunset (at altitude angles of sun below 61°) (Muneer et al., 2007). This error occurs frequently, even for very carefully calibrated pyranometers (C. Gueymard & Gueymard, 1993), and is therefore widely recognized in literature (Muneer et al., 2007). In addition, thermal offsets within the measuring instruments can occur due to radiative cooling, which can become evident by negative values at night (Roesch et al., 2011).

3.3 Quality control

The difficulty of validation studies is the comparison of two uncertain data series (modeled vs. measured). Forecast errors do not only come from model errors but are also caused by measurement errors used to verify the output of models. Since several sources of uncertainty exist, quality control (QC), as a major prerequisite for using meteorological information, can be of great benefit (Alani et al., 2021). Validation of meteorological data ensures properly generated information and identification of lacking or missing values (Estévez et al., 2011). That ensures, that the data is of satisfying reliability for the purpose of the upcoming validation. Since QC is a substantial research topic in the case of radiometric organizations, several guidelines have been published. The BSRN, for example, proposed an automated QC, whose methodologies have been widely used in literature, which this thesis will explain more in detail later on (Alani et al., 2021).

Common procedures for QC include visual inspection of meteorological observation through mapping time ranges allowing the search for missing or abnormal values or issues in time reference (Alani et al., 2021). Further investigation deals with extreme values. The so-called "range test" verifies, whether measured data lays within an upper or a lower threshold to be considered valid (Estévez et al., 2011). Within solar radiation research, certain boundaries for *GHI* are set. For example, *GHI* cannot exceed the solar constant, and, climatologically seen, cannot go below the value of zero (Journée & Bertrand, 2011). The solar constant (S_0) is defined by the radiation that reaches the top of the atmosphere at mean earth-sun distance. In literature,

it varies from 1361 W/m² to 1367 W/m² (Badescu, 2008). Physical possible limits (*PPL*) and extremely rare limits (*ERL*) , established through recommendations by the BSRN, depend on the solar constant as well as the zenith angle (*ZA*) of the sun and determine its boundaries. *PPL* and *ERL* are defined as followed (Long & Dutton, 2010):

$$PPL: \ -4 \ \leq GHI \ \leq S_0 * 1.5 * cos(ZA)^{1.2} + 100 \tag{1}$$

$$ERL: \ -2 \leq GHI \leq S_0 * 1.2 * cos(ZA)^{1.2} + 50 \tag{2}$$

The minimum limits are tolerant of the measurement errors, that can occur during night (Section 3.2). Values exceeding these limits are considered impossible and are therefore removed (Younes, 2006). When filtering measurements according to their quality, many tests rely on calculations of the solar position. Since measuring these parameters is unpractical, they are determined through calculations. In literature, no standard methodology for the latter is recommended (Perez-Astudillo et al., 2019). However, for example, the NOAA established algorithms calculating different solar parameters (NOAA, 2021). Statistic programs, such as RStudio, make use of this information and implement them in different packages (e.g. "maptools", "insol") that are easily available for its users (Bivand, 2021; Corripo, 2021). Furthermore, the behaviour of solar radiation depends on stochastic parameters like the frequency and height of the clouds, atmospheric aerosols, groundwater vapour, and atmospheric turbidity (Badescu, 2008). To isolate these stochastic components, and to focus on main error sources, the global clearness index (K_t) can be calculated. K_t is defined as the ratio of GHI and the extraterrestrial radiation (G_{ext}).

$$K_t \ = \frac{GHI}{G_{ext}} \tag{3}$$

G_{ext} itself can be calculated through a function of S_0, the sun's elevation angle (*EV*) and the Earth's orbit eccentricity correction factor (ε). The latter is calculated with a simplified equation of Spencer's eqation using the Julian day (j) as an input:

$$\varepsilon \ = \ 1 + 0.0342 \ * \ cos\left[\frac{2\pi(j-1)}{365}\right] \tag{4}$$

Accordingly, G_{ext} is calculated through the following equation (Paulescu et al., 2021):

$$G_{ext} \ = \ S_0 * \varepsilon * sin(EV) \tag{5}$$

The K_t-index ranges from zero to one. Any data exceeding these boundaries are unrealistic. Thereby, the K_t-index acts as another indicator for erroneous measured data (Younes, 2006). Finally, the knowledge of the clear sky irradiance reaching the ground is a key parameter in the field of solar radiation modeling and evaluation (Journée & Bertrand, 2011). To investigate the consistency of measured data regarding clear sky conditions, clear sky irradiations (G_{clear}) can be assessed through models. In literature, a variety of clear sky models are available. Much effort has been made to compare different models investigating advantages and disadvantages (C. A. Gueymard, 2012; Ineichen, 2016; Reno et al., 2012). In general, the options of models vary from very simple to more complicated formulations. All clear sky models require geometric inputs describing the $\measuredangle A$. Whereas the simplest model requires only the $\measuredangle A$, other simple models include further basic parameters describing the state of the atmosphere such as air pressure, temperature as well as further parameters. More complex models consider various measurable parameters like ozone, aerosols, and perceptible water. Even though being considered the most accurate models, they are usually very time-consuming in processing and also heavily dependent on local measurements. Very simple clear sky models tend to significantly underpredict the irradiance, especially for high altitude sights. In contrast, simple models accounting for the altitude are more comparable to the accuracy of more complex models. For instance, the model established by Ineichen performed almost as well as the more complicated model "REST2" (Reno et al., 2012). A more recent study by Yang (2020) supports previous assumptions as well. By comparing three different clear sky models, including the Ineichen-Perez model, Yang found out, that there is "[…] no evidence on which to base the belief that high-performance clear-sky models are superior to the simpler ones in forecasting applications". Under the circumstances, that complex clear sky models entail, the author suggests the Ineichen-Perez model being the most suitable (Yang, 2020). Ineichen and Perez added corrections to the original formulation by Kasten to the following equation (Ineichen & Perez, 2002):

$$G_{clear} = a_1 * G_{ext}^{(-a_2 * AM * (f_{h1} + f_{h2} - (T_L - 1)))} \tag{6}$$

where:

$$a_1 = 5.09 * 10^{-5} * h * 0.868 \tag{7}$$

$$a_2 = 3.92 * 10^{-5} * h * 0.0387 \tag{8}$$

$$f_{h1} = exp\left(\frac{-h}{8000}\right) \tag{9}$$

$$f_{h2} = exp\left(\frac{-h}{1250}\right) \tag{10}$$

The corrected version mainly uses the Linke Turbidity (T_L), the Air Mass (AM) and the elevation above sea level (h) of the location as input (Ineichen & Perez, 2002). T_L quantifies the atmospheric visibilty under clear sky (Journée & Bertrand, 2011). It is a convenient approximation to model the atmospheric absorption and scattering of radiation (Reno et al., 2012) and can express the optical thickness of a cloudless atmosphere (Ineichen & Perez, 2002). The more attenuation of radiation by the atmosphere, the larger the T_L (Reno et al., 2012). To create a worldwide database for T_L, Remund et al. (2008) calculated and produced maps for monthly or yearly values using a combination of ground measurements and satellite data (Remund et al., 2008). AM is a parameter that measures the path length that solar rays follow in the atmosphere before reaching the ground. The longer the path, the stronger the interaction between solar radiation and the atmospheric constituents (Badescu, 2008). The path is dependent on the $\measuredangle Z$. The higher the $\measuredangle Z$, the higher AM. At $\measuredangle Z = 90°$, the AM is one. Usually, the air mass is often approximated for a constant density atmosphere and ignores the Earth's curvature using the geometry of a parallel plate. That is why at zenith angles larger than 80° the accuracy degrades rapidly, where AM goes to infinity. A simple approximation of the AM, developed by Kasten and Young, is defined as followed (Reno et al., 2012):

$$AM = \frac{1}{cos(ZA) + 0.50572 * (96.07995 - ZA)^{-1.6354}} \tag{11}$$

When working with modeled G_{clear}, certain conditions need to be considered, since previously described calculations become imprecise when it comes to high $\measuredangle Z$'s. Furthermore, with very high zenith angles, GHI can, in fact, be higher than the G_{ext} due to the diffusive effects of clouds (Journée & Bertrand, 2011). Therefore, reasonable limits of radiation data are considered to range between $75° < \measuredangle Z < 85°$ (Alani et al., 2021; Engerer & Mills, 2015; C. A. Gueymard & Ruiz-Arias, 2016; Muneer et al., 2007; Reno et al., 2012; Yang, 2020; Younes et al., 2005). Since GHI may exceed G_{clear} occasionally due to cloud enhancement (Yang et al., 2018), the upper limit for GHI is set to 1.1 times the G_{clear} (Journée & Bertrand, 2011).

3.4 Model verification

Through verification, the accuracy of the forecast can be investigated (DWD, 2021b). Comparing the performance of different models gives valuable information not only to

researchers for further model development, but also to forecast users as an assisting tool for choosing between different forecasting products (Perez et al., 2013). To evaluate the performance of models, several statistical indices accompanied by their strengths and weaknesses are available (Behar et al., 2015). The following statistical mediums are rated as suitable in literature (Huang & Thatcher, 2017; Mathiesen & Kleissl, 2011; Mayer & Butler, 1993; Troccoli & Morcrette, 2014; Wang et al., 2012):

When working with different datasets consisting of a high amount of data points, calculating their arithmetic mean is a common approach to make generalized statements of different dataset. With *x* representing each data point, and *n* being the amount of data points, the arithmetic mean is calculated as followed (Dormann, 2017):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{12}$$

Based on the simple arithmetic mean, the mean absolute error (*MAE*) describes particularly the amount of deviation of two different datasets, for instance the prediction (*y*) from the observation (*x*) throughout all datapoints (*n*), but no direction of deviation. That means, positive and negative deviations equal out (Mayer & Butler, 1993).

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - x_i| \tag{13}$$

To assess if a model over- or underpredicts the forecast, mean bias error (*MBE*) is calculated.

$$MBE = \frac{1}{n}\sum_{i=1}^{n} y_i - x_i \tag{14}$$

Whenever the output is negative, the model tends to underpredict and vice-versa for positive output (Younes et al., 2005). To compare the interrelationships of different datasets, Pearson's correlation coefficient is used since it is a normalized measurement of the covariance measuring the coherence of different datasets. It is defined as followed (Dormann, 2017.):

$$cor(y,x) = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}} \tag{15}$$

Its value ranges from minus one to one. Positive values imply that both datasets vary in the same direction, negative values mean the smaller one dataset, the larger the other. Values around the value of zero indicate no correlation.

4 Hypotheses and objectives

The goal of this thesis is to assess the model accuracy of global weather forecast models and to optimize the forecast accuracy by a multi-model approach in order to achieve the lowest possible forecast error. Analyses will concentrate on the following initial assumptions:

It is assumed, that the accuracy of NWP-models is subject to constant investigation, and therefore, has improved within the last years [H1]. Also, combining the output of different individual models can lower the forecast error, since inaccuracies can to some degree outweigh each other [H2]. Within multi-models, it is expected, that the more models are included in the mix, the lower the forecast error will be [H3]. Furthermore, it is assumed that raw models perform unequally well for certain spatial or climatic conditions [H4]. Where seasonal and regional differences occur, effects on local multimodel performances are possible [H5].

5 Materials and methods

5.1 Data basis

The basis of this thesis was measured and modeled global radiation. Data from BSRN were downloaded for the years 2018 and 2019 and unfolded as well as converted to text files, with the help of the BSRN-Toolbox (WRMC-BSRN, 2021). Data from WRDC were copied and saved in Excel-files for the years 2018 and 2020 (WRDC, 2021). The data from the Global Atmosphere Watch (GAW) stations were used being available in an hourly resolution. Since data at WRDC are saved as tables per month, yearly time series had to be produced afterwards. In addition to radiation measurements, both networks provide the user with information about the metadata of each station. The coordinates (latitude and longitude) of each station, as well as the elevation, were separately saved for further analysis. The stations were plotted on a world map to check the plausibility of the station's location with the aid of the statistical program RStudio. Table 1 gives a summary of the data sources. Figure 1 shows the spatial resolution of the stations per dataset.

Table 1: Summary of properties of datasets used in this thesis.

	BSRN	**WRDC (GAW)**
Temporal resolution	Minute	Hour
Aggregation	Instantaneous	Backwards
Unit	W/m^2	J/cm^2
# of stations	43	43
Timezone	UTC	Local
Years	2018, 2019	2018, 2020

Since data were available in different temporal resolutions, the processing of these data for an overall consistency was mandatory. With the help of RStudio, data were loaded and further processed to consistent time series. Minute instantaneous data were averaged to hourly backward data. Each timestamp began on January 1st, 1:00 and ended on December 31st, 23:00. For every station, singular time series were produced. Furthermore, data were converted, if necessary, to the unit W/m^2. In addition, WRDC data had to be transformed from local to UTC time. Data of stations located in Australia had to be interpolated to full hours after the time zone transformation since the time-offset to UTC is nine hours and 30 minutes.

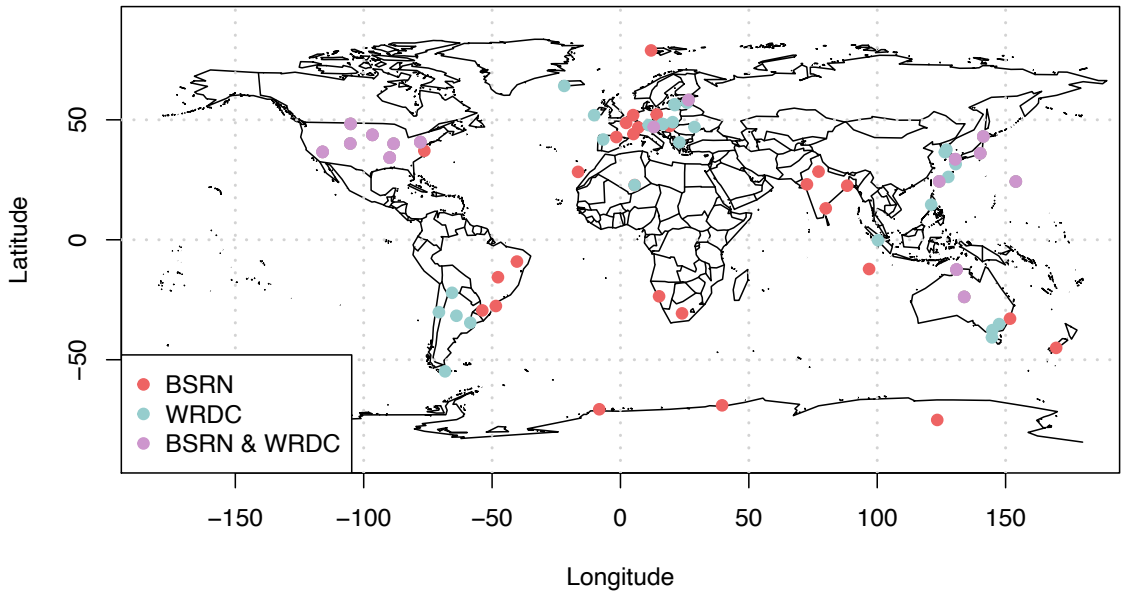


Figure 1: Spatial resolution of stations from BSRN and WRDC. A total of 16 stations are overlapping.

Daily forecast datasets (12-35 hours ahead) for the years 2018 until 2020 have been provided by meteoblue AG (meteoblue AG, 2021a). Following models were taken into consideration for solar radiation validation:

- ERA5 (2018-2020)
- GFS (2018-2020)
- NEMSGLOBAL (2018-2020)
- MFGLOBAL (2019-2020)
- ICON (2018-2020)

Note, that MFGLOBAL was not available for 2018. Backwards forecast data were downloaded for every station in an hourly time resolution.

5.2 Preparing data

Before the validation of the raw models, quality control filters were applied to the measured data. To prepare the quality control process, several parameters were defined, and specific solar data were calculated. The solar position was derived from the timestamp, latitude, longitude, and elevation of each station with the help of the RPackage "insol" (Corripo, 2021). $\mathcal{Z}A$ and EV were calculated for every minute of the required year (2018 until 2020). Afterwards, G_{ext} was calculated with Equation (4) and (5). Furthermore, after deriving AM (Eq. (11)), instantaneous G_{clear} was calculated (Eq. (6 to 10)). T_L was available as a GEOTIFF-map in a yearly resolution and downloaded online via Solar radiation Data (SoDa) (SoDa, 2010). Subsequently, G_{ext}, G_{clear}, $\mathcal{Z}A$ and EV were averaged to hourly data by calculating the arithmetic means.

5.3 Quality control procedure

Before removing or correcting erroneous data, visual inspection of the measured data was carried out. Within that, data gaps were detected, and examination of time reference issues was possible. Sunset and sunrise were calculated with the help of a function included in the RPackage "maptools" (Bivand, 2021) using the latitude and the longitude of the station as an input. For every station that showed errors, a "peak correction" was conducted. To do so, daily radiation peaks of the measured data, usually occurring around noon (local time), were visually compared to those of calculated G_{ext}. Within that, the measured data was alternated by pre- or postponing it by zero to two timesteps. For each option, Pearson's correlation coefficient was calculated (Eq. 14). The option that brought out the highest coefficient was chosen to correct the timestamp of the measured data, assuming that the peaks of both datasets lie on top of each other.

The next procedures aimed to detect extreme outliers that exceed certain physically possible and extremely rare limits. Before analyzing these, one important condition was met. As previously explained (Section 3.2), the cosine error of the measurement's instruments is a reoccurring error. To bypass these problems, all values of the measured data during sunrise and sunset were removed (Filter I). That included all values (< 0 W/m^2) where $\measuredangle A > 85°$. This ensured on the one hand, that problematic data did not lead to consequential errors later, and on the other hand, that nighttime values were still considered for further validation. Next, *PPL* (Filter II) and *ERL* (Filter III) for every hour of each day of the year were calculated based on the formulations established by the BSRN (Eq. (1 and 2)). Data were analyzed and removed as soon as it exceeded its limit. Figure 2.b shows, how visual inspection of the latter can be realized. Furthermore, due to climatological limits, all data points lower than zero were deleted (Filter IV). Thereafter, more advanced quality control followed. Further limits for measured data were set by evaluating, whether measurements exceeded clear sky radiation. In the next step, G_{clear} and G_{ext} were used to calculate the K-indices. *GHI* and G_{clear} were both divided by the G_{ext}.

$$K_t = \frac{GHI}{G_{ext}} \tag{16}$$

$$K_{t,clear} = \frac{G_{clear}}{G_{ext}} \tag{17}$$

K_t was plotted against the *EV* angle to evaluate its distribution. $K_{t,clear}$ was added (Figure 2.c). The upper limit was set to the following:

$$K_t < 1.1K_{t,clear} \quad (18)$$

All K_t points exceeding the climatological possible limit were considered erroneous and therefore deleted (Filter V). Checking the timestamp visually once more, the measured data should then lay within the sunrise and the sunset and should be free of outliers (Figure 2.a). During the QC procedure, the loss of the measured data after each filter was documented and analyzed.

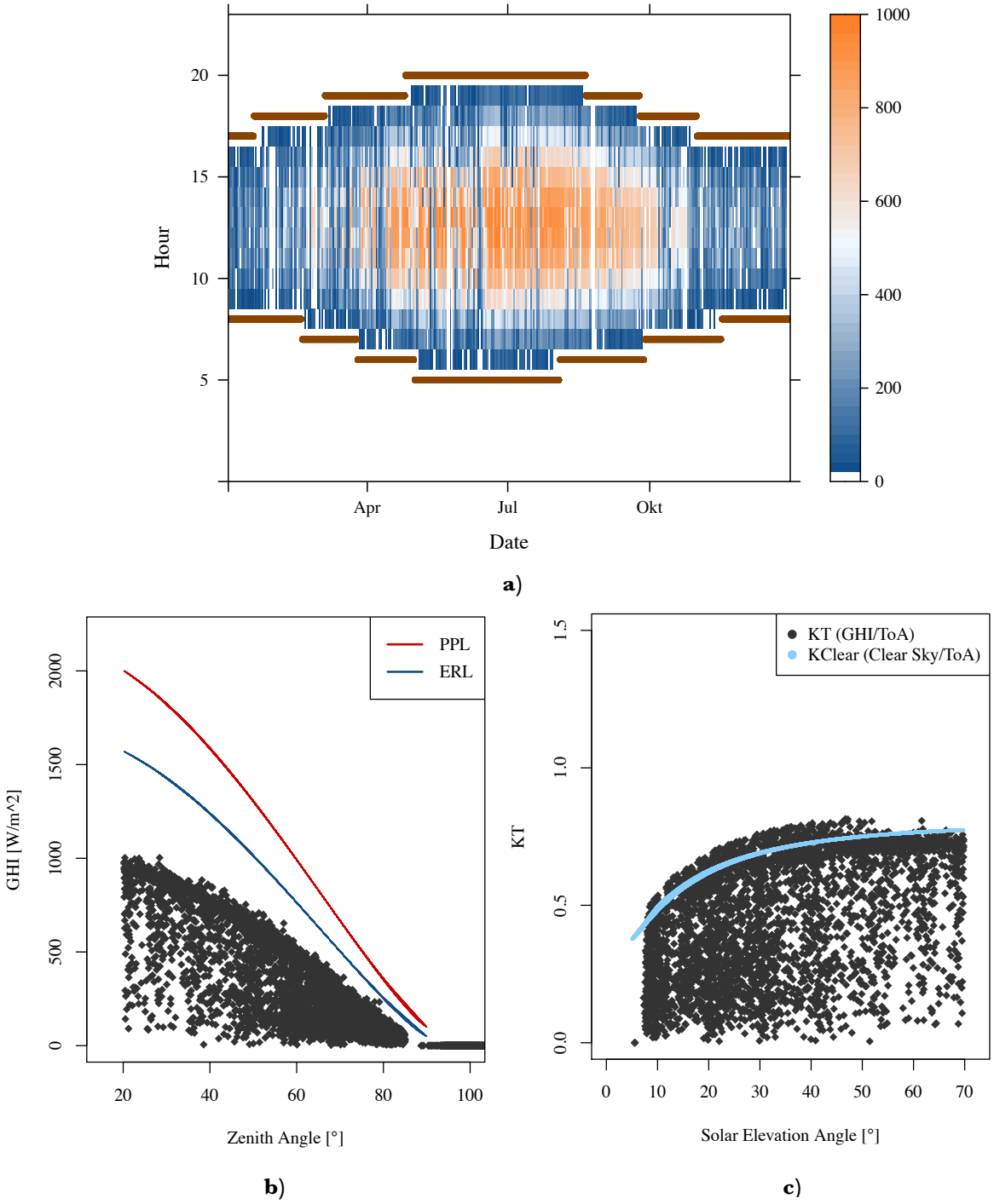


Figure 2: Visual inspection of QC-procedures, illustrated by station "CNR" (BSRN 2018). Plot a) was used to examine time consistency between measured data and solar position. All filters were applied, and peak-correction conducted. Plot b) shows, if measured solar irradiance exceeds *PPL's* or *ERL's* after applying Filters II and III. Plot c) represents the K_t-index plotted against the *EV*, after applying Filter V.

5.4 Model Verification

To evaluate models, measured data of each dataset were compared to forecast data. Statistical analysis was conducted to evaluate and compare the performance of each (raw or multi-) model by calculating the *MAE* and the *MBE*. While *MBE* and *MAE* calculate average values for each individual station, most of the results of this thesis are based on the further calculation of average values over all available stations, describing the central tendency of the data.

5.4.1 Multi-model verification

For the multi-model validation, four models were considered in total: GFS, ICON, MFGLOBAL, NEMSGLOBAL. Multi-models were formed by adding up a chosen number of models. Each model was assigned a specific weight, which was implemented as a factor.

$$MM = a * Model_1 + b * Model_2 + c * Model_3 + d * Model_4 \quad (19)$$

The models were weighted in 10 % steps. That means factors *a*, *b*, c, and *d* could range from zero to one in steps of 0,1. All factors had to add up to one (100 %). In total, there were 258 combinations available. For each combination, the *MAE* and *MBE* were calculated.
Further analysis was conducted through different approaches. First, raw model validation was conducted, where *MAE* and *MBE* of each model were compared (Section 6.2). Then, multi-model analyses followed. In Section 6.3, the differences between multi-models consisting of either two (M2), three (M3) or four (M4) raw models were examined. Scatterplots, in which the *MAE* and the *MBE* for each combination are plotted against each other, were used to examine different patterns between the combinations. After that, the combination achieving the lowest *MAE* per station was documented and used for further analysis. To point out first tendencies, the influence of the number of models within a multi-model was investigated. Within that, the frequency of how often either raw models, 2M-, 3M-, or 4-M-combinations achieved the lowest *MAE* was illustrated. Furthermore, the *MAE* of all raw models, as well as all 2M-, 3M-, 4M-combinations were separately averaged for each station. Afterwards, the mean for all stations was calculated, giving a summary of how these different combinations perform on average. In Section 6.4, all raw models were compared to the multi-models by calculating the difference of their *MAE's*. To follow up on previous analyses, it was examined, how much weight was assigned to the models to achieve the lowest *MAE*. These results were illustrated through histograms. Getting closer to a first conclusion on the best performing combination, mean percentages for each model were averaged over all combinations that achieved the lowest *MAE*. The latter was performed on M2, M3 and M4 separately (Section 6.5).

To find out if the multi-model approach is robust against seasonal variabilities, all possible combinations were performed for each quarter of the year. Therefore, measured and modeled data were split every third month in order to represent each season. For each station and quarter, the 20 best combinations achieving the lowest *MAE's* were filtered out. All quarters were compared by determining the number of overlapping combinations, under the assumption, that more overlaps suggest higher robustness of the multi-model approach. Thereafter, the means of all combinations for every quarter were calculated and compared to each other (Section 6.6).

In Section 6.7, data were simply plotted on world maps to investigate the abundance of spatial patterns. First, for each dataset and year, the number of overlapping combinations in steps of two was plotted on a world map and tagged with different colors. The second part of the spatial analysis consisted of plotting the percentage of each model, through which analysis of spatial differences on where models are weighted higher and lower were possible. Here, the combinations of the best *MAE* per station were used. Last, the *MAE* per station of each model were plotted on world maps to examine spatial similarities of raw model performances and individual model weightings.

6 Results

6.1 Quality controlled data

Table 2 shows the means of every single filter, averaged over all stations within one dataset and one year. The first column shows the overall percentage of data gaps that already existed before the quality control. Data gaps, originated by station errors or unreliable data assimilation vary strongly throughout the stations. On average, 3-25 % of yearly data were missing. In general, the data from WRDC showed fewer gaps than the data from BSRN. Most gaps occurred in the data of the BSRN in 2019 when on average up to 1/4th of each station was missing. In contrast, the data of WRDC 2018, showed few missing data.

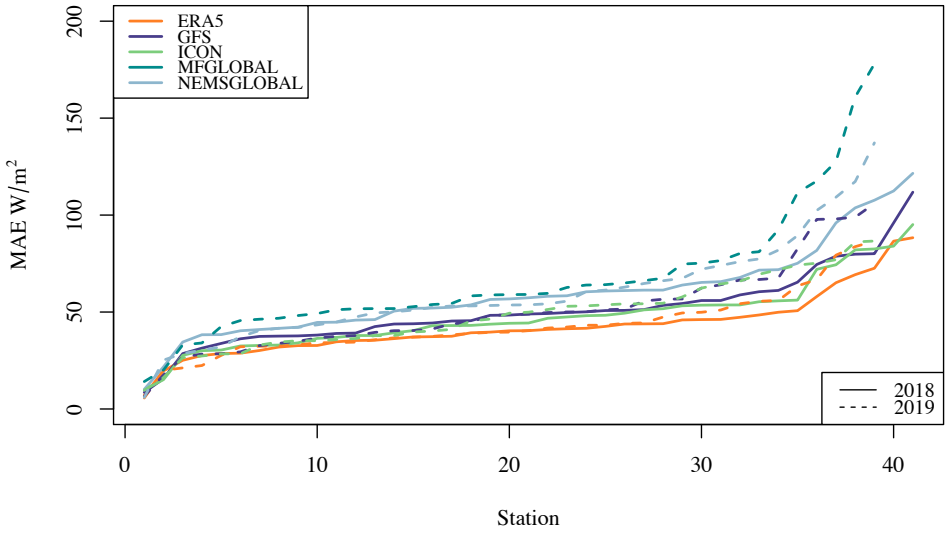
Table 2: Quantity [%] of data points affected by each filter. All filters deleted data except for Filter IV. Here, data were instead corrected to zero, therefore not included in "Removed".

	Data gaps	**ZA > 85°**	**Min PPL**	**Max PPL**	**Min ERL**	**Max ERL**	**Nega-tives**	**Clear sky**	**Re-moved**
# Filter		I	II	II	III	III	IV	V	I-III, V
BSRN 18	10.30	10.45	1.60	0.00	4.26	0.00	12.84	5.31	21.62
BSRN 19	24.97	8.98	1.92	0.00	4.30	0.01	12.17	4.89	20.09
WRDC 18	2.27	6.89	0.00	0.00	0.00	0.13	0.00	5.34	12.45
WRDC 20	10.22	6.44	0.00	0.00	0.00	0.07	0.00	4.60	11.10

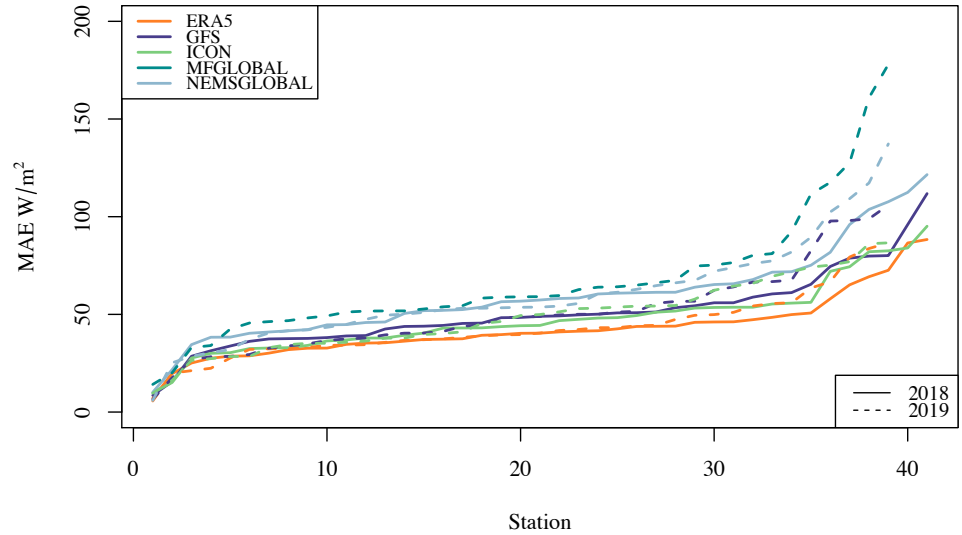
Looking at the other QC-filters, most of the data were lost through Filter I, deleting the sunset and sunrise data (6-11 %). Only a few values (up to 2 % per station) exceeded the *PPL's* and *ERL's*. Most of them included the negative night values that arose from the nighttime offset of the measuring instruments. On average, no measurement contained outliers exceeding the maximum *PPL*, only very rarely (WRDC 2018, 2020) measurements exceeded the maximum *ERL*. The remaining negative values (up to 13 % on average) were set to zero. Only data from BSRN showed negative nighttime offsets. For the clear sky filter, approximately 5 % of the data exceeded the clear-sky-limitation and were removed. All in all, more data from the BSRN had to be removed, than from the WRDC dataset.

6.2 Raw model verification

Figure 3 and Table 3 summarize the results of the raw model validation. In Figure 3, the performance of each model is represented by showing the *MAE* per station. In this figure, a ranking of the model-performances is visible. Corresponding mean values for each dataset and year are enclosed in the appendix (Table 12). Both datasets show the following results. The lowest *MAE* were seen in the ERA5 reanalysis model, with a maximum *MAE*-average of 44.01 W/m^2. The second-best model is ICON (max. average 49.57 W/m^2). GFS (max. average 51.92 W/m^2) and NEMSGLOBAL (max. average 61.08 W/m^2) follow. The highest *MAE* were calculated for the model MFGLOBAL, with an average *MAE* up to 67.39 W/m^2, especially shown for the stations of BSRN. On average, the *MAE* throughout all



a) BSRN



b) WRDC

Figure 3: Performance of each model illustrated by the *MAE* per station for both datasets separately.

models varied from 27.28 W/m² to 67.39 W/m². Table 3 shows the mean values for the *MAE* and the *MBE* for every model averaged over all same stations within one dataset. That allowed comparison between the years of one dataset. After analyzing the *MBE*, it was shown that MFGLOBAL and NEMSGLOBAL underestimated the measured data (negative values), while the other models overestimated the measured data. In general, the best *MBE* was met with ICON for WRDC 2020 (1.03 W/m²). When comparing the errors of different years, with one exception (ICON), the *MAE's* declined from 2018 to 2019 and 2018 to 2020. The *MAE* of ERA5 decreased by up to 2.11 W/m², of GFS by up to 2.26 W/m², and of NEMSGLOBAL by up to 2.1 W/m². For the average *MBE*, improvements varied. The average *MBE* of ICON and GFS (for both datasets) improved, whereas, for example, the average *MBE* of ERA5 for the measurements of BSRN deteriorated.

Table 3: Mean absolute error [W/m²] (left) and mean bias error [W/m²] (right) averaged over the stations of each dataset, that were available for 2018 and 2019. In 2018, MFGLOBAL was not available (n.a).

	BSRN		**WRDC**		**BSRN**		**WRDC**	
	2018	**2019**	**2018**	**2020**	**2018**	**2019**	**2018**	**2020**
	MAE				*MBE*			
ERA5	43.27	42.61	39.56	37.45	6.23	8.02	8.70	6.33
ICON	48.05	48.35	42.54	42.99	13.76	8.31	14.48	1.03
GFS	51.13	49.82	46.80	44.54	13.88	12.35	19.66	13.24
NEMSGL.	60.62	59.61	52.76	50.61	-9.0	-8.62	-2.60	-4.64
MFGL.	n.a	67.14	n.a	51.56	n.a	-42.86	n.a	-25.59

6.3 Multi-model verification and analyses

6.3.1 Comparison of *MAE* and *MBE*

The performance of the raw models, as well as the multi-models, could be illustrated with scatterplots for each station. Figure 4 shows scatterplots of selected stations. The different colored dots represent the number of models used in each multi-model. The stations 1-4 showed noticeable results, in which the combination with the lowest *MAE* achieved a low *MBE* as well. Furthermore, it can be seen, that all multi-model combinations achieved a better *MBE* within the highest and the lowest *MBE* calculated by the raw models. Counterexamples to the just mentioned stations showed the stations e and f. In very rare cases there was a possibility, in which no multi-model combination could lower the *MAE*, meaning an individual model performed better (station e). Station f showed that low *MAE's* do not always mean low *MBE's*.

In general, these scatterplots also show the distribution of different combinations. While the data points describing for the 2M-combinations are variably distributed and therefore more spread out, the data points representing the 4M-combinations, are far more bundled and less scattered.

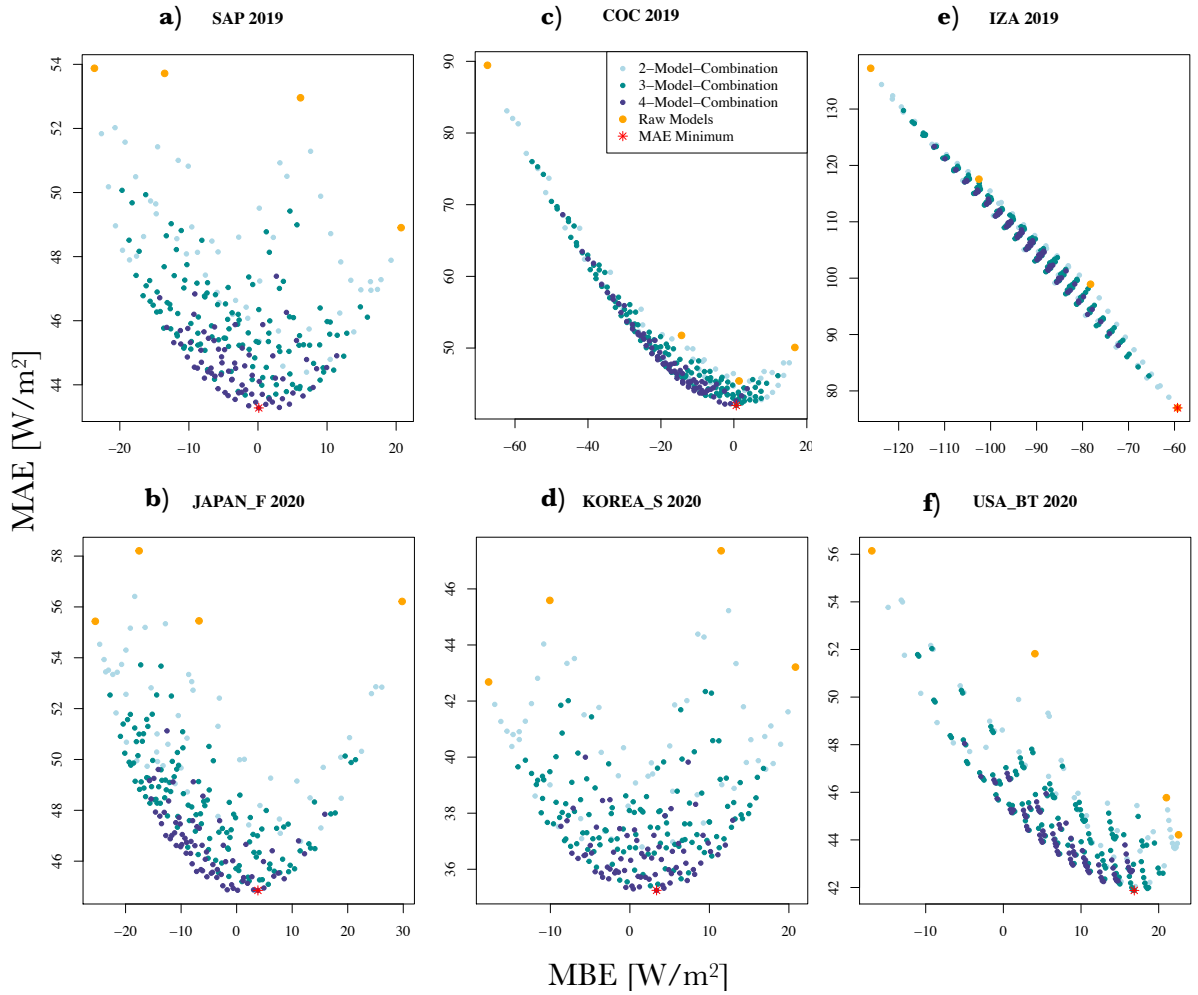


Figure 4: Scatterplots of *MAE* and *MBE* of all multi-model combinations. On the left (a and b) and in the middle (c and d), stations are shown where M4-combinations achieve the lowest *MAE*. The best combinations as well show low *MBE's*. On the right (e and f), stations are illustrated which form exemptions. For IZA 2019, no multi-model could lower the *MAE*. USA_BT shows, that low *MAE's* might as well mean unsatisfying *MBE's*.

6.3.2 Influence of the number of models within a multi-model mix

In the following, this thesis examined, how often different multi-model combinations achieve the lowest *MAE*. In Figure 5, the frequency of the best multi-models regarding their number of raw models is illustrated. The blue bar represents the multi-model combination that obtained the lowest *MAE* the most recurrently within a year throughout all stations of one dataset. For 2018, there were no model mixes of four models available. For up to 40 stations, the 3M-combination achieved the lowest *MAE* within WRDC and BSRN. For one station of the

dataset WRDC, the raw model performed better than any other combination. For 2019 and 2020, 4M-combinations were available. The lowest *MAE's* for these years were achieved by 4M-combinations. In 2019, almost the same number of 3M-combinations as 4M-combinations achieved the best results. Even though two models combined or even the raw models themselves can produce good results for some stations, the lowest *MAE's* however tend to be achieved when using three or four models. Examining, how much more the *MAE* is lowered by adding another model to the mix gave a more detailed insight into how the different variations perform.
Results are summarized in the Table 4. The biggest and most significant decrease of *MAE* was achieved, when two models were combined and compared with the raw models themselves. Thereby, the *MAE* was lowered by up to 4.5 W/m^2. When a third model was added to the model mix, *MAE* decreased again but was less significant. The error was maximally reduced by 2.4 W/m^2. Using four models, *MAE* was lowered again, however, by even less (approximately 1.7 W/m^2).

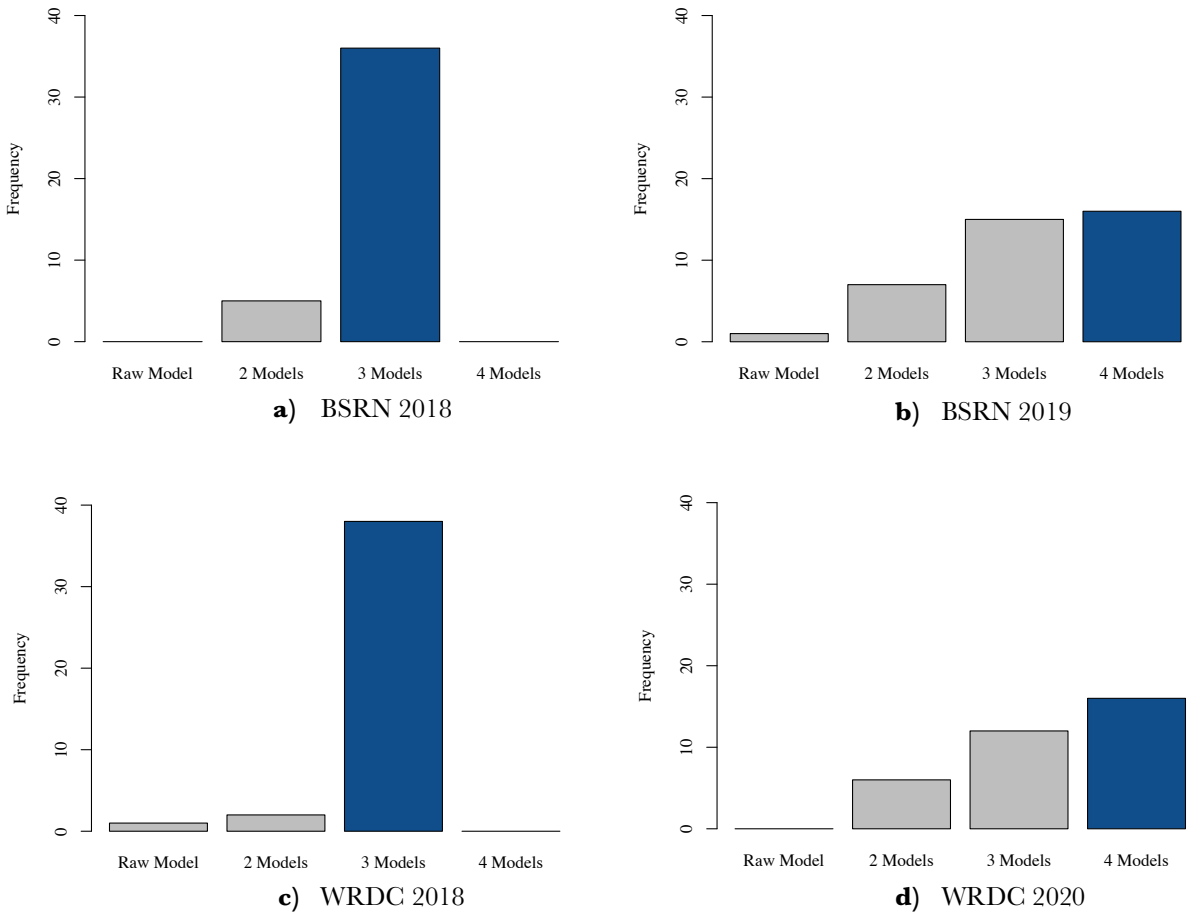


Figure 5: Barplots showing how often raw models, M2, M3 or M4-combinations achieved the lowest *MAE* per dataset and year.

Table 4: *MAE* [W/m²] averaged over all stations per dataset and year for raw models, 2M-, 3M, 4M-combinations.

	Raw Models	2 Models	3 Models	4 Models
BSRN 2018	53.27	48.93	46.72	n.a
BSRN 2019	56.44	51.79	49.37	47.64
WRDC 2018	49.46	45.64	43.62	n.a
WRDC 2020	47.66	43.97	42.08	40.79

6.4 Comparison of multi-models with raw models

To examine the impact of multi-model mixes, the statistical error was compared to those of the raw models. It was shown in Section 6.2, that ERA5 had on average the best *MAE's*. In the next step, it was investigated, if multi-models lower the *MAE* even below the best raw model (that was, as a reminder, not included in the multi-models), and it was shown how much the error in comparison to all raw models has been lowered.

Table 5 shows the average mean (best) *MAE* values of each dataset's stations and all years of first, the multi-model, and second, every raw model. In addition, the difference of the multi-model to each individual raw model is shown. A negative prefix means a lower *MAE*, a positive therefore a larger *MAE*. First, the multi-model was compared to ERA5. Except for BSRN 2018, there was always a decrease of the *MAE* identifiable through the multi-model identifiable. The

Table 5: Comparison of the *MAE* [W/m²] of multi-model (MM) results (blue) with all raw models. One exemption is marked in bold.

	BSRN 2018	BSRN 2019	WRDC 2018	WRDC 2020
MM	43.85	41.17	40.81	37.26
ERA5	42.69	42.97	41.48	37.28
DIFF	**+1.16**	-1.8	-0.67	-0.02
GFS	50.80	50.74	48.82	44.48
DIFF	-6.95	-9.57	-8.01	-7.22
ICON	49.34	48.70	45.97	42.84
DIFF	-5.49	-7.53	-5.17	-5.59
MFGLOBAL	n.a	66.85	n.a	52.25
DIFF	n.a	-25.68	n.a	-14.99
NEMSGLOBAL	59.67	59.46	53.60	51.07
DIFF	-15.82	-18.30	-12.79	-13.81

MAE has been reduced by up to 1.8 W/m^2. Yet, the *MAE's* were still very similar to those of the ERA5 model. Comparing the *MAE's* of the multi-models to all other raw models, there was always a decrease of the *MAE* visible. These results were much more significant than the comparison of the multi-model to ERA5. Looking at NEMSGLOBAL for example, the *MAE* could have been lowered by up to 18.3 W/m^2. The *MAE* has decreased by almost 30 %. The results for MFGLOBAL showed even higher decreases. Where MFGLOBAL had an error of 66.85 W/m^2 for BSRN (2019), the multi-model lowered the error to 41.17 W/m^2. That was a decrease of 40 % of the raw model error. All in all, the multi-model lowered the *MAE* compared to the raw models (GFS, ICON, MFGLOBAL, NEMSGLOBAL). On average, the multi-model approach reduced the *MAE* by 6.95 W/m^2 to 25.86 W/m^2.

6.5 The best multi-model-combinations

To find the best multi-model combination, this thesis investigated the weighting of each model for the best *MAE* per station. Hereby, a general overview was given by the histograms shown for every dataset and year (Figure 6). A general pattern is recognizable. Looking at the bar plots for ICON, a usual distribution can be seen around 0.5, especially in the year 2018. While ICON was weighted the highest most of the time, GFS and NEMSGLOBAL were added to similar components. In 2019 and 2020, where 4M-combinations were available, several models were frequently not considered in the multi-model mix (frequency of 0 % is relatively high). In summary, the weightings of 2019 showed similar tendencies to 2018. In 2020, there was a variable weighting of the model ICON, and higher percentages for NEMSGLOBAL and GFS more often. In general, the distributions of the models' ratios were more balanced than in the year 2018. Specifically looking at MFGLOBAL, it was used rarely for the combinations. Yet, for a small number of stations, it was weighted very high.

Previous results are summarized in Table 6. The best performing combinations were

Table 6: Weighting of each model after the best combinations were averaged over all stations per dataset.

	GFS	**ICON**	**MFGLOBAL**	**NEMSGLOBAL**
BSRN 2018	0.33	0.43	n.a	0.24
WRDC 2018	0.28	0.48	n.a	0.25
BSRN 2019	0.24	0.39	0.17	0.21
WRDC 2020	0.21	0.46	0.14	0.19

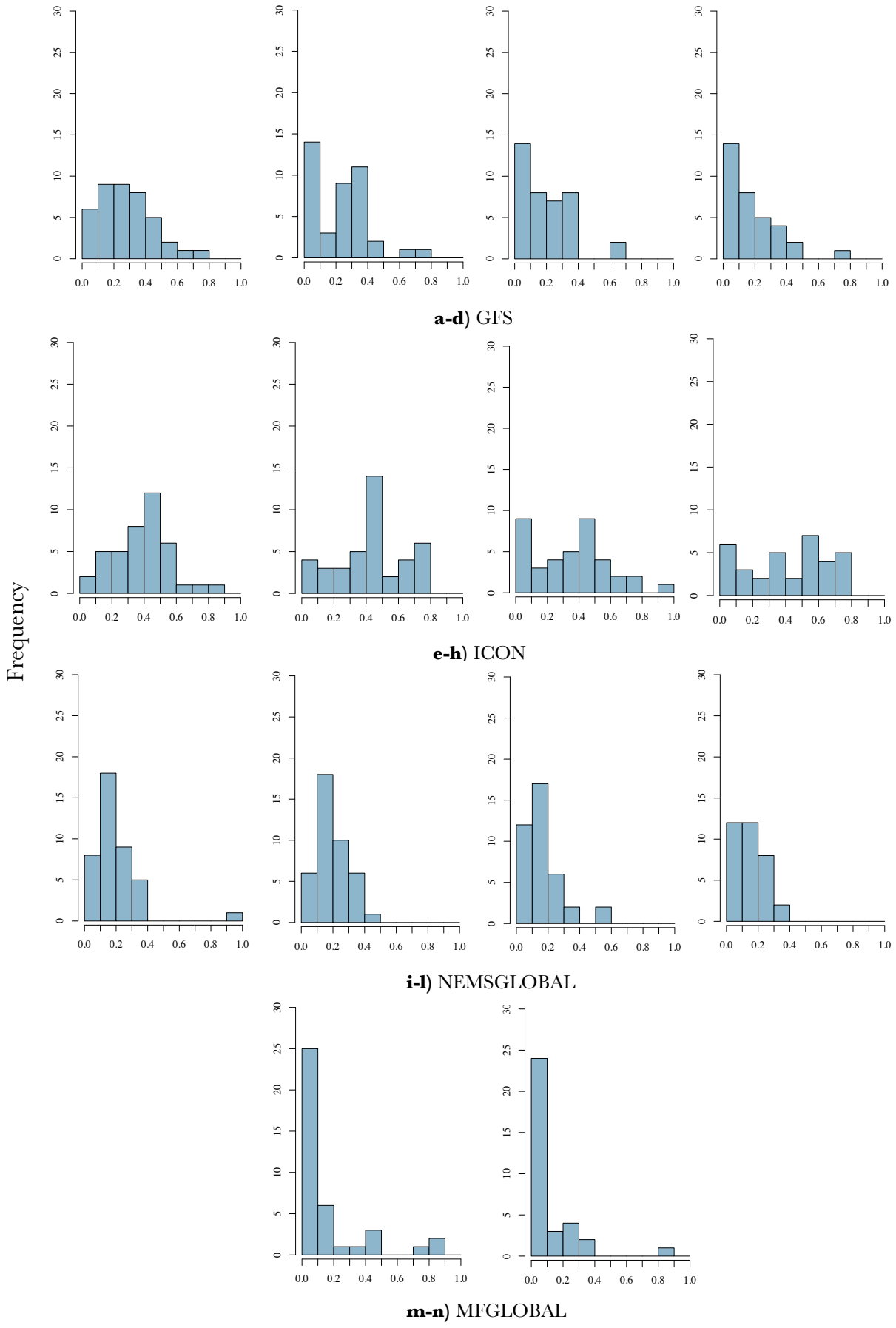


Figure 6: Histograms showing how each model was weighted for each station to achieve the lowest *MAE*. The x-axis describes, how high the models were weighted, the y-axis represents the number of stations affected. From left to right, BSRN 2018, BSRN 2019, WRDC 2018 and WRDC 2020 are shown. M) and n) are an exemption and show only results for BSRN 2018 (m) and WRDC 2020 (n).

averaged over every station within one dataset. It shows the general best distribution of models within a multi-model for each dataset and year. All in all, as previously described, ICON was weighted the highest, followed by GFS and NEMSGLOBAL. In 2019 and 2020, MFGLOBAL was added, but weighted only a small percentage (up to 17 %). Yet, even though it performed quite poorly as a raw model, it still lowered the *MAE* when combined with other models. For 2018, a model combination of three models could lower the error. The combination, that would bring on average the best results for both datasets can approximately be described as:

$$0.3 * GFS + 0.5 * ICON + 0.2 * NEMSGLOBAL \quad \text{(Ia)}$$

For 2019 and 2020, considering a 4M-combination, the best multi-model can approximately be described as:

$$0.2 * GFS + 0.4 * ICON + 0.2 * MFGLOBAL + 0.2 * NEMSGLOBAL \quad \text{(IIa)}$$

Note, that factors were rounded to the 10th since initially the combinations were calculated in steps of 10.

Furthermore, the means of the best combinations with model mixes consisting of either two, three or four models were calculated. With these results, it was possible to see, which models are preferably used within 2M- and 3M-combinations. Within that, all models were considered available. That means for example, by investigating the lowest *MAE's*, the best 2M-combination out of all available raw models was able to be chosen. Table 7 shows the most frequent best performing 2M-combinations and 3M-combinations, that occurred throughout all stations. Looking at 2M-combinations first, it becomes clear that the combination ICON = 70 % and NEMSGLOBAL = 30 % was the most frequent combination within all samples. Even in 2019 and 2020, where a fourth raw model was available, ICON and NEMSGLOBAL performed the best the most often. For 3M-combinations, the results varied. In general, the combination of GFS, ICON and NEMSGLOBAL performed the best. Throughout the weightings, there are slight differences within GFS and NEMSGLOBAL. ICON commonly took up 50 % of the multi-model, except in 2020, when it was weighted even higher (80 %). Table 8 shows the 4M-combinations. All models were considered, explaining why no results for the year 2018 were available. Particularly in 2019, there were inconsistencies between the model weightings. While ICON was weighted the highest in the first two combinations, MFGLOBAL was weighted more than 50 % in the third example. Besides this one exception, it was recognizable, that, once again, ICON was weighted the highest generally, whereas GFS and NEMSGLOBAL took up less weight. MFGLOBAL was being weighted even lower (10 %). The previous results give a clue on how to weigh different models whenever using a 2M-, a 3M- or a 4M-mix.

Table 7: Combinations, that achieved the lowest *MAE*, shown for 2M- and 3M-combinations. For all stations, the 2M- (3M-)combinations achieving the lowest *MAE* were chosen, and compared to each other. The most common best performing combinations within all stations per dataset and year are summarized in this table.

	2-combination		Frequency	**3-combination**			Frequency
Models	ICON	NEMS-GLOBAL		GFS	ICON	NEMS-GLOBAL	
BSRN 2018	0.7	0.3	24 %	0.3	0.5	0.2	15 %
BSRN 2019	0.7	0.3	13 %	0.4	0.5	0.1	8 %
				0.3	0.5	0.2	8 %
WRDC 2018	0.7	0.3	12 %	0.4	0.5	0.1	7 %
				0.3	0.5	0.2	7 %
WRDC 2020	0.7	0.3	21 %	0.1	0.8	0.1	15 %

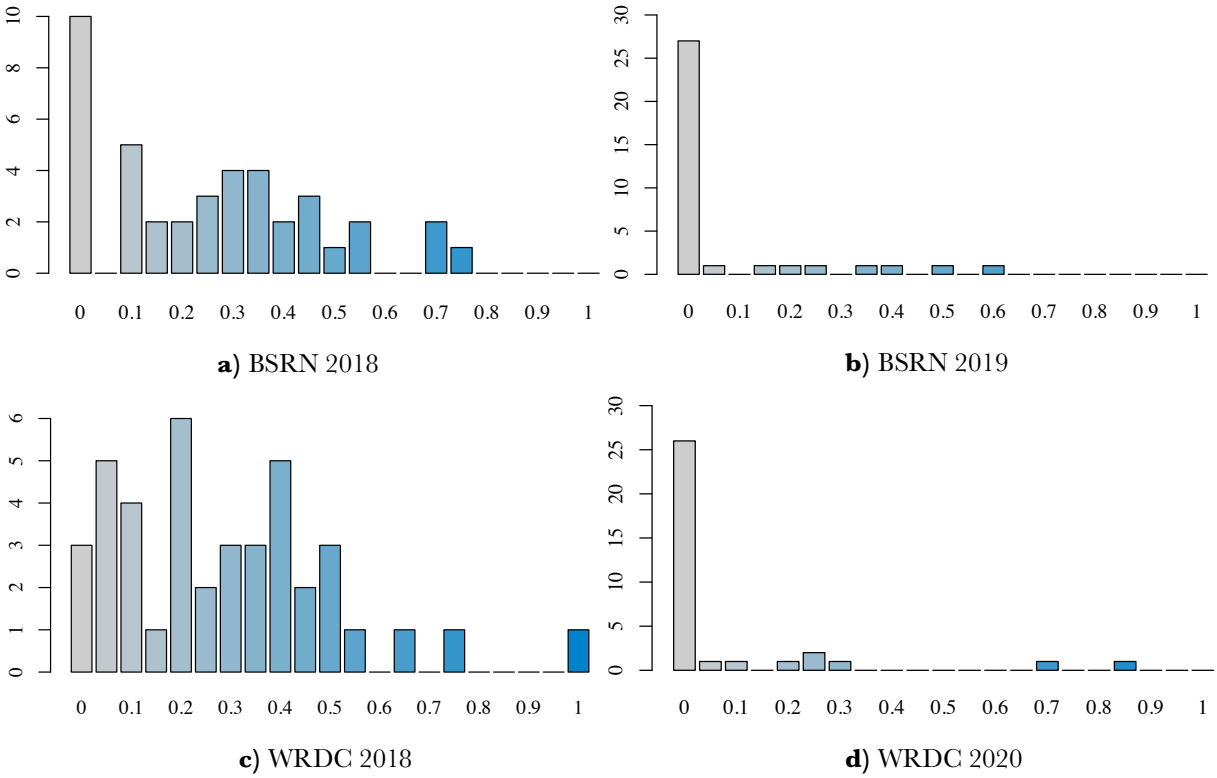
Table 8: Combinations that achieved the lowest *MAE*, shown for 4M-combinations. For all stations, the 4M-combinations achieving the lowest *MAE* were chosen and compared to each other. The most common combinations within all stations per dataset and year are summarized in this table.

	4-combination				Frequency
Models	GFS	ICON	MFGLOBAL	NEMSGL	
BSRN 2019	0.3	0.5	0.1	0.1	8 %
	0.3	0.4	0.1	0.2	8 %
	0.1	0.1	0.6	0.2	8 %
WRDC 2020	0.2	0.6	0.1	0.1	15 %

6.6 Seasonal robustness

To show, if the multi-model approach is robust within a seasonal variability, this thesis analyzed the quarterly differences of multi-model combinations and their *MAE's* within a year. The first approach was to check, if there were overlapping multi-model combinations within the 20 best *MAE's* of each quarter.

For each dataset and each year, Figure 7 shows bar plots of the frequency of overlapping combinations between the four quarters of each station, represented by an index from zero to one. The more overlapping combinations, the higher the index. An index of one means, that all 20 best combinations of each quarter were the same. For BSRN 2018, 10 of 41 (22 %)



a) BSRN 2018 **b**) BSRN 2019

c) WRDC 2018 **d**) WRDC 2020

Figure 7: The number of overlapping combinations within the 20 best *MAE's* of each quarter of every station. The x-axis represents an index describing the relative amount of overlapping combinations out of all 20 best combinations within all four seasons of one station. The y-axis shows, how many stations are affected. Note, that the resolution of the y-axes between the four bar plots differs.

stations did not have any overlapping combinations. The remaining stations had a varying number of overlapping combinations, namely from one to 15. WRDC 2018 had only three out of 41 (7 %) stations with no overlapping combinations. The number of stations affected by indices from 0.2 to 0.5 is relatively high, and one station showed an index of 1. For BSRN 2019, there were more than 27 of 39 (70 %) stations that had no overlapping combinations. Only a few had overlapping stations, with an index varying from 0.05 to 0.6. WRDC 2020 showed similar results, where 26 out of 34 (76 %) stations had no overlapping combinations. The remaining stations had one to six overlaps ($0.05 < \text{index} < 0.3$), whereas two stations formed the exemption by having 14 (index = 0.7) and 17 (index =0.85) overlapping combinations. In 2018, where 3M-combinations were compared, many overlapping combinations were found. In 2019 and 2020, only a few overlapping 4M-combinations existed.

For comparing the means of the 20 best combinations between each quarter, the tables for chosen stations are shown in Table 9. Several stations, for instance CAB and LATVIA_Z, showed similar results within each quarter, whereas others (ASP and ARGENTINA_LQ) showed less similarity and the combinations seemed to be more random. Within station ARGENTINA_LQ, especially the first and the fourth quarter differed significantly from the second and third quarter, mostly seen for NEMSGLOBAL. Stations with less continuity within the quarters tended to have fewer or non-matching combinations, as was analyzed in the previous section. Therefore, stations with more consistency between each quarter had more overlapping combinations.

Table 9: Means of the 20 best combinations per quarter, demonstrated by four examples.

	CAB 2018				**LATVIA_Z 2020**			
quarter	GFS	ICON	MFGL.	NEMS-GL.	GFS	ICON	MFGL	NEMS-GL.
1st	0.25	0.54	n.a	0.21	0.09	0.15	0.74	0.03
2nd	0.20	0.50	n.a	0.30	0.00	0.11	0.76	0.05
3rd	0.17	0.56	n.a	0.27	0.13	0.13	0.72	0.02
4th	0.24	0.58	n.a	0.16	0.09	0.15	0.74	0.03
	ASP 2018				**ARGENTINA_LQ 2020**			
1st	0.64	0.12	n.a	0.25	0.35	0.59	0.02	0.04
2nd	0.09	0.34	n.a	0.58	0.15	0.18	0.00	0.68
3rd	0.18	0.25	n.a	0.68	0.11	0.15	0.02	0.72
4th	0.26	0.58	n.a	0.16	0.54	0.42	0.02	0.03

By calculating the mean of every quarter (Table 10) for all stations, an average and overall consistency of the combinations became visible. Looking at BSRN 2018, the maximum difference between each quarter within each model was only up to 13 %, whereas for the years 2019 and 2020 this difference increased (2019: 16 % and 2020: 21 %). Looking at the mean of all quarters for every model, an averaged combination was calculated. The results for BSRN and WRDC in 2018 seemed to be consistent with each other since they did not differ much.

Summarized, to receive a seasonally robust combination for the year 2018, including all available models (in total three), the models were combined as followed:

$$0.3 * GFS + 0.4 * ICON + 0.3 * NEMSGLOBAL \quad \text{(Ib)}$$

Looking at the year 2019 and 2020, less of GFS, ICON and NEMSGLOBAL was weighted in the multi-model since MFGLOBAL was considered in the multi-mix. All in all, the means of the quarters were less consistent, than in 2018. However, the best multi-model, being on average seasonally robust throughout all stations for 2019 and 2020, would consist of:

$$0.2 * GFS + 0.4 * ICON + 0.2 * MFGLOBAL + 0.2 * NEMSGLOBAL \quad \text{(IIb)}$$

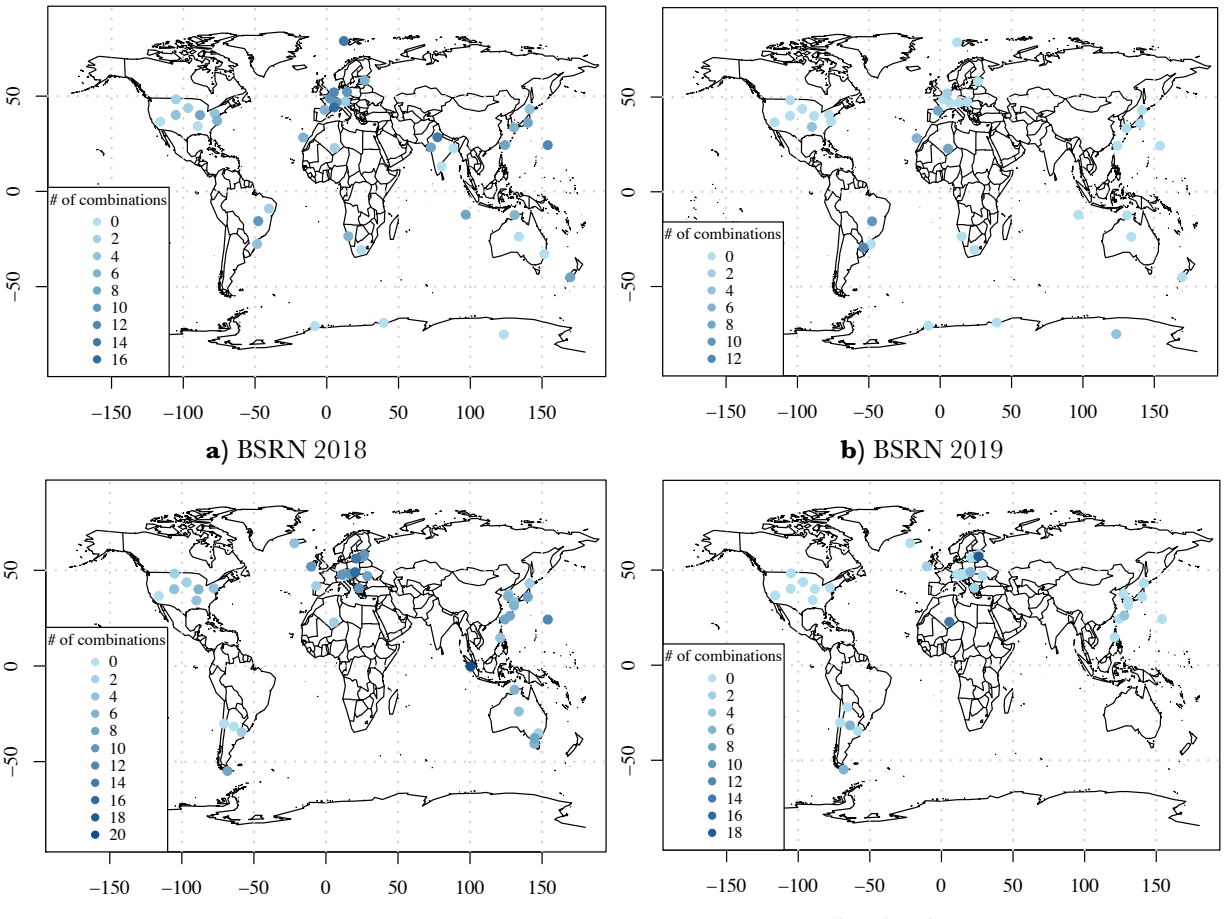
Table 10: Means of the 20 best combination per quarter averaged over all stations per dataset.

	BSRN 2018				**WRDC 2018**			
quarter	GFS	ICON	MFGL.	NEMS-GL.	GFS	ICON	MFGL	NEMS-GL.
1st	0.39	0.34	n.a	0.27	0.35	0.37	n.a	0.29
2nd	0.31	0.43	n.a	0.27	0.26	0.47	n.a	0.27
3rd	0.30	0.44	n.a	0.27	0.30	0.45	n.a	0.26
4th	0.30	0.47	n.a	0.24	0.27	0.50	n.a	0.24
	BSRN 2019				**WRDC 2020**			
1st	0.18	0.37	0.25	0.20	0.17	0.35	0.30	0.19
2nd	0.26	0.38	0.09	0.27	0.27	0.37	0.11	0.26
3rd	0.32	0.34	0.09	0.26	0.28	0.35	0.12	0.25
4th	0.21	0.35	0.24	0.21	0.14	0.43	0.32	0.13

6.7 Spatial analysis

For a more accurate forecast, it is advantageous to examine different spatial patterns of the multi-model combinations that were investigated previously. As examined in Section 6.6, the different extent of consistency between the quarters within different stations gave clue that for some spatial conditions the multimodal approach could be more suitable than for others. In

addition, it is interesting to look at spatial patterns when it comes to the weighting of different models. Regions might exist, where one raw model is typically weighted higher or lower than in other regions. It is useful to find these patterns to apply that knowledge in operational procedures and to choose the best multi-model under certain conditions. In this thesis, the spatial distribution of the stations and their number of overlapping combinations were investigated. Vague patterns are identifiable in Figure 8. The darker the color of the data point, the more overlapping combinations exist. In 2018, there were up to 20 overlapping combinations. The stations prone to more consistency within the quarters could be found in Europe and in East Asia. Therefrom affected were Japan, Korea, and the Philippines. In addition, fewer overlapping combinations were available in the stations throughout North America, Australia, and Antarctica. The most striking value of 20 overlapping stations was found in Indonesia (WRDC 2018). When looking at the following years, fewer patterns were evident because the number of overlaying multi-model-mixes was overall small. In 2019, overlapping combinations were found within the stations of North Africa and South America. In 2020, one noticeable station found in Latvia reached 17 overlapping combinations.

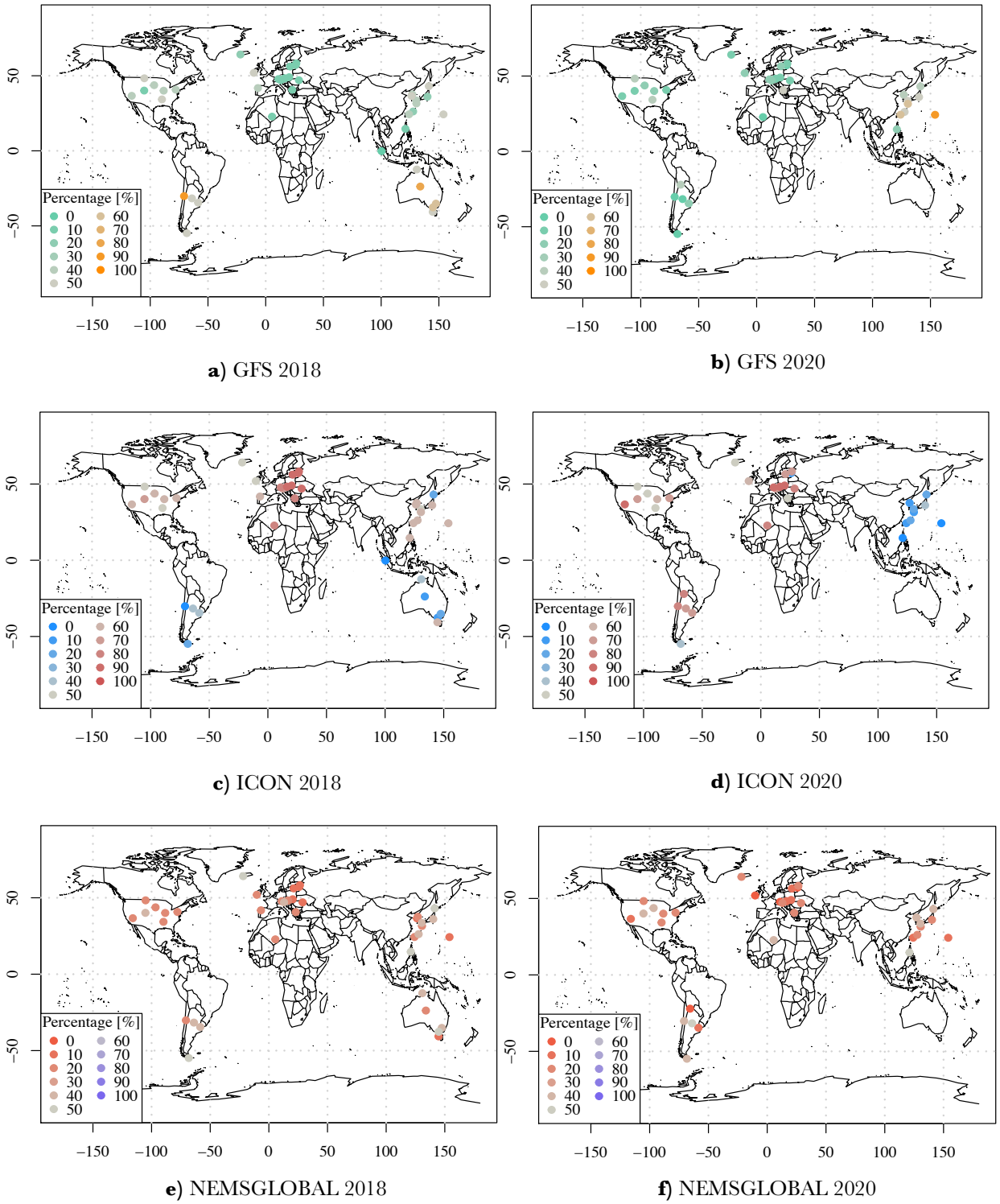


a) BSRN 2018 **b**) BSRN 2019

c) WRDC 2018 **d**) WRDC 2020

Figure 8: Spatial distribution of the number of overlapping combinations of each quarter's 20 best performing combinations per station for each dataset and year. The x-axis represents the longitude, the y-axis the latitude.

The second part of the spatial analyses examined the distribution of the weightings of the different models. For visualization, the combinations of the best *MAE* per station were used. The legend helps to understand, where the model has been weighted high or low. In Figure 9, only the stations of WRDC are used to illustrate upcoming results, since similar results can be seen for BSRN. First, GFS was examined. Low weightings (0-20 %) could be observed in



a) GFS 2018 **b**) GFS 2020

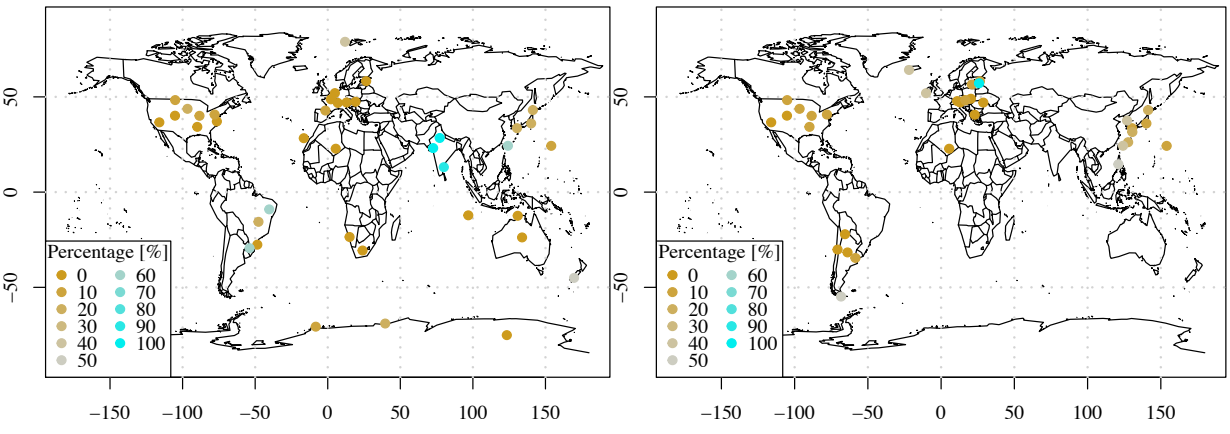
c) ICON 2018 **d**) ICON 2020

e) NEMSGLOBAL 2018 **f**) NEMSGLOBAL 2020

Figure 9: Spatial distribution of the weighting of chosen models within the stations of WRDC.

Europe and North Africa. Higher weightings (up to 80 % or more) were seen in America, especially in 2018. For 2020 GFS was weighted less, especially in South America. In East Asia, GFS was weighted moderately in 2018 (10-50 %), whereas in 2020 it had more weight (up to 90 %). More striking patterns were identifiable, when looking at the spatial distribution of the weightings of ICON. In America, approximately 50 % of ICON were used in the multi-models, albeit the tendency in the North was higher than in the South. In Europe, as well as Africa (North), ICON was highly weighted (up to 90 %). In contrast, low usage was identifiable in Australia, the Antarctic, (South) East Asia (Japan, Indonesia, Philippines, Korea). Here, the colors show a vigorous contrast.

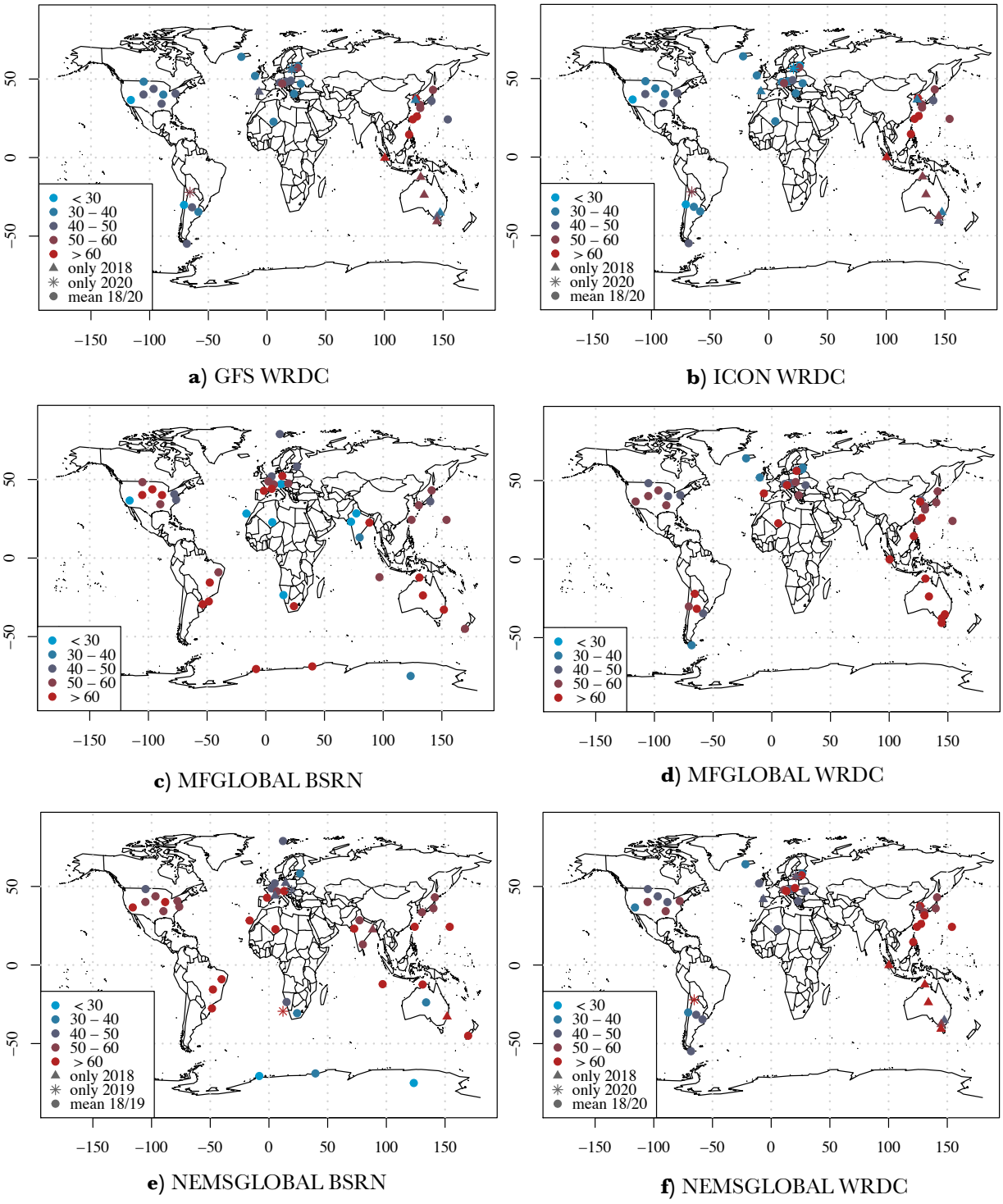
Spatial patterns regarding MFGLOBAL, illustrated in Figure 10, could only be identified for the years 2019 and 2020. Low weightings could be observed in Africa, North and South America, Australia, and Europe (< 40 %). Several stations in East Asia and in New Zealand had in contrast to the other mentioned areas relatively high MFGLOBAL-weightings, in particular in 2020 (40-60 %). Remarkably high weightings were found in India, where MFGLOBAL took up approximately 90 % of the multi-model, as well as in one station in Europe (LATVIA_Z). NEMSGLOBAL was weighted similarly in almost all regions, where stations were available, varying from 20 % to 50 %). The following figure illustrates the regional performance showing the *MAE's* per station of each individual model for both datasets. When both years were available for a particular station, the mean *MAE* was calculated. Different data points showed stations only available for one particular year. Results show similar patterns to Figure 9. GFS had rather high *MAE's* in East Asia and Australia, whereas moderate *MAE's* were found in Europe. Slightly lower *MAE's* were seen in America. ICON had high *MAE's* especially in East Asia and Australia (up to more than 60 W/m^2) and moderate *MAE's* in South America (up to 40 W/m^2), North America and Europe. Two exceptional stations in Europe exist, where the



a) MFGLOBAL 2019 **b)** MFGLOBAL 2020

Figure 10: Spatial distribution of the weighting of MFGLOBAL within the stations of BSRN (2019) and WRDC (2020).

MAE is 50-60 W/m^2. MFGLOBAL and NEMSGLOBAL had rather high *MAE's* (over 60 W/m^2) throughout all stations, whereas MFGLOBAL performed very well in India with *MAE's* lower than 30 W/m^2 as well as for some stations in Africa. NEMSGLOBAL had notably low *MAE's* in Antarctica, South Africa, and rather moderate to high *MAE's* in Europe.



a) GFS WRDC

b) ICON WRDC

c) MFGLOBAL BSRN

d) MFGLOBAL WRDC

e) NEMSGLOBAL BSRN

f) NEMSGLOBAL WRDC

Figure 11: Spatial distribution of the performance of each model, represented by the MAE [W/m^2].

7 Discussion

When interpreting analyses conducted in this thesis, it must be kept in mind, that the results of the multi-model analyses account specifically for these four models. Additional models could change the results suggesting an even more precise multi-model mixture. Important to note is also, that the multi-models were optimized by the *MAE*. As the results of Section 6.3 show, it is certain that multi-models do not optimize every error in the same way. The choice of a different statistical error, such as *MBE* or root mean squared error (*RMSE*), through which extreme errors are weighted even more, might suggest different combinations. In some literature, different statistical errors are even combined considering their different strengths and weaknesses (Behar et al., 2015; Huang et al., 2018; Yagli et al., 2019). Nonetheless, in literature, the *MAE* is a widely recognized measure for the performance of the models. Since the goal of this bachelor thesis was to give an insight into multi-models and to show their potential of improving the forecast, the *MAE* was considered as fully adequate to achieve this goal.

The results of the quality control make it possible to roughly estimate the quality of the measured data. For example, where high amounts of data gaps could be observed, the impression of partially unsatisfying quality was given. However, high gaps arose from a rather small number of stations with a substantial percentage of data missing. Data gaps varied strongly, where stations were embossed by either very small or very high data losses, with the former predominating. The errors seemed to be very random and to depend on the location, since similar data gaps could be observed for stations located in the same country (see appendix, Table 11). Since this thesis' analyses were mainly based on average values of all stations, these rare, however large amounts of missing data might have considerably influenced or even hindered the deriving of reliable means or trends within examined data. Individual stations could therefore have had a great impact. Table 3 underlines that since the omission of several stations could considerably influence the average, through which important information is extracted (such as the realization, that raw models underwent an improvement).

By applying Filter I, data was deleted and new data gaps were produced. Nevertheless, it must be beared in mind, that Filter I removed data during sunrise and sunset. For the purposes of the analyses, however, these data points had less relevance because very little radiation is present at these times. Also, only small amounts of extreme outliers were present. It is obvious, that BSRN data are strongly characterized by values undercutting the lower limits created by the thermal offsets of the stations. In contrast, WRDC data had no such errors, but were sometimes measured too high and therefore exceeded upper boundaries of extremely rare limits. The stations of WRDC seem to be well-calibrated for the thermal offsets. However, there are

references in literature questioning, that current calibration approaches can compensate directly for these errors (Badescu, 2008).

Furthermore, all stations possessed several data points exceeding the clear sky radiation. Reasons for this are mostly related to *DIF*, which, under unstable cloud conditions, may exceed maximum limits (Alani et al., 2021). In cases like these, it can be beneficial to investigate not only *GHI*, but also its components to conduct tests based on their internal consistency. In fact, state-of-the-art instrumentation for global radiation suggests a combination of a pyrheliometer and a shaded pyranometer. Therefore, global irradiance is recommended to be calculated as the sum of its components, rather than measured directly (C. A. Gueymard, 2008), thereby possibly specifying QC procedures and model verifications. Yet, since the analysis of *DIF* and *DIR* would have gone beyond the scope of this thesis, it was not considered. Quality analyses like these show the need for further reduction of extreme errors to prevent convoluted conclusions.

When analyzing the performance of raw models, ERA5 operated particularly well. Nevertheless, it was excluded intentionally from the multi-models since it is a reanalysis model and therefore cannot be equated with the others. When looking at the raw models' performances, they over- or underpredicted the radiation. For solar radiation forecasts, the prediction of the development of clouds is specifically important. When models overestimate radiation, they tend to underpredict cloud schemes and vice-versa (Tuononen et al., 2019). ERA5 is a good example showing the advantage of combining model outputs with additionally derived climatic data and therefore improving the forecast. Furthermore, results give an indication to the improvement of raw models within the last years. Yet, for more consistency within the time series analyzed in this thesis, data of three consecutive years (2018, 2019 and 2020) for all stations (from BSRN and WRDC) might have been able to show more significant results.

The multi-model analysis gave interesting results. When comparing the *MAE* and the *MBE* of different multi-model combinations, various conclusions could be deduced from the distribution of the data points. Some 2M-combinations might as well achieve low *MAE's* and *MBE's* for a particular station but are followed by a high variability and uncertainty throughout the different weightings of the models. In general, the use of M2-combinations might achieve, to some extent, lower *MAE's* but seems to be very sensitive to the weightings, possibly lowering the forecast security for each station. Looking at the model combinations with more than two models, less distribution around low *MAE's* could be observed, leading to the assumption that the certainty of a more accurate forecast increases when using M3- and M4-combinations. All in all, results of the multi-model analyses show that mixing several raw models has the potential to lower the

MAE, even outperforming the best raw model (ERA5). Certain multi-models were able to lower the error up to a significant 40 % of the raw model's output. Results also show that the more models are used, the lower the forecast error will be. However, the outcome of Table 4 strengthens the assumption that the adding of models into the multi-model reaches saturation at some point. That demonstrates that including more and more models in the operation might not be worth the expenses for a relatively little decrease of the forecast error. Hence, this thesis examined the best combinations for different numbers of raw models used in the multi-model.
Considering the availability for up to three models in the year 2018, and up to four models in the years 2019 and 2020, this thesis investigated the optimal average combination for all stations. Several patterns are recognizable. ICON was always weighted the highest, particularly in 2018, when only 3M-combinations were available. In 2019 and 2020, ICON was weighted minimally less, which could, on the one hand, be due to the availability of a fourth raw model, and on the other hand, due to no improvement of the raw model. In contrast, for GFS and NEMSGLOBAL, an improvement throughout the years was recognized, which might explain the higher weightings of these models. During 2019 and 2020, there are several 3M-combinations that performed well, which explains the high amount of non-used models (percentage = 0 %) seen in Figure 5. Furthermore, the weighting of the models seems to be less variable in 2018 than in 2019 and 2020, which could be related to the 4th raw model. With adding MFGLOBAL, more options possibly lowering the error arise. However, this model did not achieve as good results as the other raw models, which could be an explanation for the poor usage within the multi-models. Results of Table 7 and Table 8 underline these assumptions, showing more variability of well performing multi-model combinations when comparing 2M-, 3M-, and 4-M combinations. While the best combination for 2M-combinations was obvious, numerous combinations for 3M-, and 4M-combinations performed equally well.
Summarizing these results shows, that ICON is always part of the multi-model, since it performs rather well when it comes to the forecast error *MAE*. GFS takes up a small part, when choosing three or four models to combine, but can rather be neglected, when it comes to 2M-combinations. NEMSGLOBAL seems to perform well especially next to ICON, regardless of GFS's outperformance. Both have a low bias (Table 3). While NEMSGLOBAL tends to underpredict, ICON tends to overpredict the radiation. Combined, they seem to equal each other out to some extent. For seasonal robustness, the number of the overlapping combinations of the 20 best *MAE's* gives the first impression, that the multi-model approach is less robust than expected and seems to be variable as well as to depend on the location of the station. Even if there is no exact concordance of the best combinations, some only differ by 10 % per model. Also, merely steps of 10 % were conducted within the analyses. Steps of 5 % or even 1 % would

have given a more detailed weighting of different models within the mix. Averaging all 20 best combinations, however, leads to more promising results, especially for the year 2018. When looking at 2019 and 2020, less consistency can be observed, which underlines possible seasonal differences of different raw models expressed especially by 4M-combinations with more alternative combinations being available. When comparing the weightings of two different approaches conducted within this thesis (Section 6.5 and 6.6), it can be seen, that both approaches result in very similar or even the same combinations, which consolidates the reliability of these combinations. In general, these results show, that the multi-model approach is, on average, a robust method for solar radiation forecasting, whereas the variability of seasonal consistency throughout different stations indicates, that spatial differences for robustness of the multi-models exist. Identifying these regional differences could have a great impact on regional forecast accuracy.

While investigating spatial analyses, several facts became evident. Usually, NWPs tend to perform less well, especially in tropical climates, where convective clouds are climate determining factors. For some continents or even countries, raw models seem to perform better, than for others. Roughly speaking, NEMSGLOBAL and MFGLOBAL seem to perform globally consistent, whereas ICON and GFS perform highly variable. These variabilities can be seen when examining the spatial patterns of performance and the weightings of raw models within multi-models. Where raw models show high *MAE's* and therefore tend to perform less well, they are considered in the multi-models to a smaller extent. Where they perform well, they are preferably used in the multi-model. To give more detailed conclusions, the spatial resolution of available stations was unsatisfying, and therefore no further analyses were conducted. Nevertheless, these results indicate the existence of spatial differences. These differences might even be related to climatic conditions changing during the year, as the spatial analyses of the overlapping combinations (Section 6.7) suggest, showing regional inconsistencies of seasonal robustness. Results show the potential for further investigation, which could lead to an even more accurate forecast, including adapted weightings of models within multi-models to local conditions.

8 Conclusion and future research

This thesis has shown that multi-models consisting of multiple NWP's form a potential approach for improving solar radiation forecasts. Compared to several models investigated in this thesis, multi-models were able to lower the forecast error by up to 40 %. It was shown, that the higher the number of included raw models in the multi-model, the lower the forecast error. Within that, however, the significance of improvement decreases. Furthermore, through comparing different multi-model combinations, this thesis calculated the globally best performing mixture. Raw models having been improved within the last years show the importance of steady verification of their performance. Since verification usually relies on local measurements, quality-controlling these measurements is of major necessity to conduct reliable information about NWP's. While raw models perform differently well, especially seen in a spatial and seasonal context, differently weighing them in certain regions and times per year can lead to an even more accurate forecast. That as well shows the potential of not only the investigation of global, but also of regional NWP's to be implemented in multi-models. While already existing models are constantly improving, the availability of new models on the market is increasing. These findings suggest a perpetual investigation of different multi-model combinations that could, for instance, be realized by machine learning methods. Within these, several local climatic conditions could be used to calculate the best multi-model automatically and periodically, even on a daily basis.

Appendix

Table 11: Data gaps [%] shown for every single station and both years.

WRDC	2018	2020	BSRN	2018	2019
ALGERIA_T	0.41	59.69	**ASP**	2.40	24.90
ARGENTINA_BA	0.32	8.57	**BON**	0.27	0.22
ARGENTINA_LQ	--	0.05	**BOS**	0.26	0.13
ARGENTINA_P	4.06	1.05	**BUD**	--	41.36
ARGENTINA_U	3.74	3.71	**BRB**	46.55	67.47
AUSTRALIA_AP	0.10	--	**CAR**	67.31	--
AUSTRALIA_CG	0.74	--	**CAB**	1.53	0.08
AUSTRALIA_DA	0.10	--	**CNR**	1.38	58.63
AUSTRALIA_MA	3.01	--	**COC**	0.55	17.40
AUSTRALIA_WW	2.83	--	**DAA**	0.00	16.15
AUSTRIA_G	0.02	0.02	**DOM**	4.28	10.73
AUSTRIA_S	0.02	0.02	**DRA**	0.07	0.08
AUSTRIA_WHW	0.02	0.02	**DWN**	1.04	22.27
CHILE_ET	0.07	1.64	**FLO**	0.75	5.64
ESTONIA_TT	0.01	58.49	**FPE**	0.90	1.03
GERMANY_H	0.16	0.05	**FUA**	0.03	0.58
GREECE_T	2.97	7.07	**GAN**	91.95	91.51
ICELAND_R	0.01	0.01	**GCR**	0.29	0.24
INDONESIA_BK	9.69	--	**GOB**	1.99	0.00
IRELAND_V	0.02	0.02	**GUR**	7.35	91.51
JAPAN_F	1.58	25.87	**HOW**	61.72	--
JAPAN_I	2.03	25.57	**GVN**	9.34	1.36
JAPAN_K	0.09	25.23	**ISH**	0.35	0.01
JAPAN_M	0.29	25.72	**IZA**	0.08	0.02
JAPAN_N	0.09	25.23	**LAU**	0.46	67.40
JAPAN_S	1.39	25.57	**LIN**	0.06	--
JAPAN_T	0.13	25.23	**LRC**	0.32	0.64
KOREA_S	0.40	0.11	**MNM**	0.07	0.11
KOREA_A	7.43	--	**NEW**	32.02	--
LATVIA_L	--	1.37	**NYA**	0.01	1.46
LATVIA_R	30.59	--	**PAY**	0.42	0.01
LATVIA_Z	18.15	0.48	**PAL**	0.02	16.92
MOLDOVA_K	0.01	0.01	**PSU**	0.14	6.91
PHILIPPINES_QC	0.08	0.08	**PTR**	61.47	91.51
PORTUGAL_B	0.82	--	**SAP**	8.48	0.31
SLOVAKIA_PG	0.00	25.14	**SMS**	--	73.48
USA_B	0.17	0.08	**SON**	0.35	1.52
USA_BT	0.35	0.16	**SXF**	0.18	9.03
USA_DR	0.18	0.59	**SYO**	0.47	3.00
USA_FP	0.23	0.16	**TAM**	0.50	2.31
USA_GC	0.17	0.14	**TAT**	0.00	0.00
USA_RP	0.17	0.09	**TIR**	16.74	91.61
USA_SF	0.18	0.17	**TOR**	0.00	0.26

Table 12: *MAE* and *MBE* [W/m²] averaged over all available stations per year.

	BSRN 2018	**BSRN 2019**	**WRDC 2018**	**WRDC 2020**	**BSRN 2018**	**BSRN 2019**	**WRDC 2018**	**WRDC 2020**
	MAE				*MBE*			
ERA5	43.41	44.01	41.48	27.28	7.37	8.87	10.10	5.65
ICON	48.31	49.57	44.93	42.84	14.77	10.42	18.20	0.36
GFS	51.48	51.93	48.82	44.69	15.60	14.89	21.74	12.11
NEMSGLOBAL	60.47	61.08	53.69	51.07	-7.51	-9.21	-1.95	-5.99
MFGLOBAL	--	67.39	--	52.25	--	-41.78	--	-26.97

References

Alani, O. el, Ghennioui, H., Ghennioui, A., Saint-Drenan, Y. M., Blanc, P., Hanrieder, N., & Dahr, F. E. (2021). A visual support of standard procedures for solar radiation quality control. *International Journal of Renewable Energy Development*, *10*(3), 401–414. https://doi.org/10.14710/ijred.2021.34806

Antonio, M., Machuca, C., & Hyndman, R. (2018). *Forecasting: Principles and Practice Related papers Automatic Time Series Forecasting: the Forecast Package for R*.

Badescu, V. (2008). *Modeling Solar Radiation at the Earth Surface*.

Behar, O., Khellaf, A., & Mohammedi, K. (2015). Comparison of solar radiation models and their validation under Algerian climate - The case of direct irradiance. *Energy Conversion and Management*, *98*, 236–251. https://doi.org/10.1016/j.enconman.2015.03.067

Bivand, R. (2021, November 7). *Maptools*. https://www.rdocumentation.org/packages/maptools/versions/1.1-2

Boisserie, M., Arbogast, P., Descamps, L., Pannekoucke, O., & Raynaud, L. (2014). Estimating and diagnosing model error variances in the Météo-France global NWP model. *Quarterly Journal of the Royal Meteorological Society*, *140*(680), 846–854. https://doi.org/10.1002/qj.2173

Cantelaube, P., & Terres, J.-M. (2005). Seasonal weather forecasts for crop yield modelling in Europe. *Tellus A: Dynamic Meteorology and Oceanography*, *57*(3), 476–487. https://doi.org/10.3402/tellusa.v57i3.14669

CNRM. (2021). *ARPEGE*. https://www.umr-cnrm.fr/spip.php?article121&lang=en

Corripo, J. (2021, February 10). *RPackage "insol."* https://www.rdocumentation.org/packages/insol/versions/1.2.2

Dormann, C. (2017). *Statistik und ihre Anwendungen Parametrische Statistik* (2nd ed.). Springer Spektrum. https://doi.org/10.1007/978-3-662-54684-0

DWD. (2021a). *Globalmodell ICON*. https://www.dwd.de/DE/forschung/wettervorhersage/num_modellierung/01_num_vorhersagemodelle/icon_beschreibung.html?nn=19912

DWD. (2021b). *Numerische Vorhersagemodelle*. https://www.dwd.de/DE/forschung/wettervorhersage/num_modellierung/01_num_vorhersagemodelle/numerischevorhersagemodelle_node.html

DWD. (2021c). *Verifikation numerischer Wettervorhersage*. https://www.dwd.de/DE/forschung/wettervorhersage/num_modellierung/05_verifikation/verifikation_node.html

ECMWF. (2021a). *Climate reanalysis*. https://www.ecmwf.int/en/research/climate-reanalysis

ECMWF. (2021b, November 22). *ERA5: data documentation*. https://confluence.ecmwf.int/display/CKB/ERA5%3A+data+documentation#ERA5:datadocumentation-Introduction

Engerer, N. A., & Mills, F. P. (2015). Validating nine clear sky radiation models in Australia. *Solar Energy*, *120*, 9–24. https://doi.org/10.1016/j.solener.2015.06.044

Estévez, J., Gavilán, P., & Giráldez, J. v. (2011). Guidelines on validation procedures for meteorological data from automatic weather stations. *Journal of Hydrology*, *402*(1–2), 144–154. https://doi.org/10.1016/j.jhydrol.2011.02.031

Gregory, P. A., Rikus, L. J., & Kepert, J. D. (2012). Testing and diagnosing the ability of the bureau of meteorology's numerical weather prediction systems to support prediction of solar energy production. *Journal of Applied Meteorology and Climatology*, *51*(9), 1577–1601. https://doi.org/10.1175/JAMC-D-10-05027.1

Gueymard, C. A. (2008). REST2: High-performance solar radiation model for cloudless-sky irradiance, illuminance, and photosynthetically active radiation - Validation with a benchmark dataset. *Solar Energy*, *82*(3), 272–285. https://doi.org/10.1016/j.solener.2007.04.008

Gueymard, C. A. (2012). Clear-sky irradiance predictions for solar resource mapping and large-scale applications: Improved validation methodology and detailed performance analysis of 18 broadband radiative models. In *Solar Energy* (Vol. 86, Issue 8, pp. 2145–2169). https://doi.org/10.1016/j.solener.2011.11.011

Gueymard, C. A., & Ruiz-Arias, J. A. (2016). Extensive worldwide validation and climate sensitivity analysis of direct irradiance predictions from 1-min global irradiance. In *Solar Energy* (Vol. 128, pp. 1–30). Elsevier Ltd. https://doi.org/10.1016/j.solener.2015.10.010

Gueymard, C., & Gueymard, C. (1993). *Critical analysis and performance assessment of clear sky solar irradiance models using theoretical and measured data* (Vol. 51, Issue 2).

Heinemann, D., Lorenz, E., & Girodo, M. (2006a). *Forecasting of Solar Radiation*.

Heinemann, D., Lorenz, E., & Girodo, M. (2006b). *Solar irradiance forecasting for the management of solar energy systems*.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., … Thépaut, J. N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Huang, J., Rikus, L. J., Qin, Y., & Katzfey, J. (2018). Assessing model performance of daily solar irradiance forecasts over Australia. *Solar Energy*, *176*, 615–626. https://doi.org/10.1016/j.solener.2018.10.080

Huang, J., & Thatcher, M. (2017). Assessing the value of simulated regional weather variability in solar forecasting using numerical weather prediction. *Solar Energy*, *144*, 529–539. https://doi.org/10.1016/j.solener.2017.01.058

Ineichen, P. (2016). Validation of models that estimate the clear sky global and beam solar irradiance. *Solar Energy*, *132*, 332–344. https://doi.org/10.1016/j.solener.2016.03.017

Ineichen, P., & Perez, R. (2002). *A new airmass independent formulation for the linke turbidity coefficient* (Vol. 73, Issue 3). www.elsevier.com/locate/solener

Journée, M., & Bertrand, C. (2011). Quality control of solar radiation data within the RMIB solar measurements network. *Solar Energy*, *85*(1), 72–86. https://doi.org/10.1016/j.solener.2010.10.021

Kleissl, J. (2010). *Current State of the Art in Solar Forecasting*.

Long, C. N., & Dutton, E. G. (2010). *BSRN recommended QC tests*.

Mathiesen, P., & Kleissl, J. (2011). Evaluation of numerical weather prediction for intra-day solar forecasting in the continental United States. *Solar Energy*, *85*(5), 967–977. https://doi.org/10.1016/j.solener.2011.02.013

Mayer, D. G., & Butler, D. G. (1993). Statistical validation. In *Ecological Modelling* (Vol. 68).

meteoblue AG. (2021a). *Weather History Download*. https://www.meteoblue.com/en/historyplus#

meteoblue AG. (2021b, November 23). *meteoblue Weather-Simulation*. https://docs.meteoblue.com/en/meteo/data-sources/datasets#nmm)

Muneer, T., Younes, S., & Munawwar, S. (2007). Discourses on solar radiation modeling. In *Renewable and Sustainable Energy Reviews* (Vol. 11, Issue 4, pp. 551–602). https://doi.org/10.1016/j.rser.2005.05.006

NCEI. (2021). *Numerical Weather Prediction*. https://www.ncei.noaa.gov/products/weather-climate-models/numerical-weather-prediction

NCEP. (2021). *GFS, Global Forecast System*. https://www.dwd.de/DE/forschung/wettervorhersage/num_modellierung/01_num_vorhersagemodelle/numerischevorhersagemodelle_node.html

NOAA. (2021). *Solar Calculator Links*. https://gml.noaa.gov/grad/solcalc/sollinks.html

Paulescu, M., Paulescu, E., & Badescu, V. (2021). Nowcasting solar irradiance for effective solar power plants operation and smart grid management. In *Predictive Modelling for Energy Management and Power Systems Engineering* (pp. 249–270). Elsevier. https://doi.org/10.1016/b978-0-12-817772-3.00009-4

Perez, R., Lorenz, E., Pelland, S., Beauharnois, M., van Knowe, G., Hemker, K., Heinemann, D., Remund, J., Müller, S. C., Traunmüller, W., Steinmauer, G., Pozo, D., Ruiz-Arias, J. A., Lara-Fanego, V., Ramirez-Santigosa, L., Gaston-Romero, M., & Pomares, L. M. (2013). Comparison of numerical weather prediction solar irradiance forecasts in the US, Canada and Europe. *Solar Energy*, *94*, 305–326. https://doi.org/10.1016/j.solener.2013.05.005

Perez-Astudillo, D., Bachour, D., & Martin-Pomares, L. (2019). *Effect of Solar Position Calculations on Filtering and Analysis of Solar Radiation Measurements*. 1–9. https://doi.org/10.18086/eurosun2018.09.08

Reinert, D., Prill, F., Frank, H., Denhard, M., Baldauf, M., Schraa, C., Gebhardt, C., Marsigli, C., & Zängl, G. (2021). *DWD Database Reference for the Global and Regional ICON and ICON-EPS Forecasting System Version 2.1.7*. https://doi.org/10.5676/DWD_pub/nwv/icon_2.1.7

Remund, J., Perez, R., & Lorenz, E. (2008). *Comparison of solar radiation forecasts for the USA*. http://www.weather.gov/ndfd/

Reno, M. J., Hansen, C. W., & Stein, J. S. (2012). *Global Horizontal Irradiance Clear Sky Models: Implementation and Analysis*. http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

Roesch, A., Wild, M., Ohmura, A., Dutton, E. G., Long, C. N., & Zhang, T. (2011). Assessment of BSRN radiation records for the computation of monthly means. *Atmospheric Measurement Techniques*, *4*(2), 339–354. https://doi.org/10.5194/amt-4-339-2011

Roulston, M. S., Bolton, G. E., Kleit, A. N., & Sears-Collins, A. L. (2006). *A Laboratory Study of the Benefits of Including Uncertainty Information in Weather Forecasts*. http://www.meteo.psu.edu/roulston/wx_

SoDa. (2010). *Linke Turbidity Factor Worldwide*. http://www.soda-pro.com/help/general-knowledge/linke-turbidity-factor

Troccoli, A. (2010). Seasonal climate forecasting. In *Meteorological Applications* (Vol. 17, Issue 3, pp. 251–268). John Wiley and Sons Ltd. https://doi.org/10.1002/met.184

Troccoli, A., & Morcrette, J. J. (2014). Skill of direct solar radiation predicted by the ECMWF global atmospheric model over Australia. *Journal of Applied Meteorology and Climatology*, *53*(11), 2571–2588. https://doi.org/10.1175/JAMC-D-14-0074.1

Tuononen, M., O'Connor, E. J., & Sinclair, V. A. (2019). Evaluating solar radiation forecast uncertainty. *Atmospheric Chemistry and Physics*, *19*(3), 1985–2000. https://doi.org/10.5194/acp-19-1985-2019

Wang, F., Mi, Z., Su, S., & Zhao, H. (2012). Short-term solar irradiance forecasting model based on artificial neural network using statistical feature parameters. *Energies*, *5*(5), 1355–1370. https://doi.org/10.3390/en5051355

WRDC. (2021). *WRDC registration*. http://wrdc.mgo.rssi.ru/wrdc_en_new.htm

WRMC-BSRN. (2021). *Data retrieval via ftp*. https://bsrn.awi.de/?id=387

Yagli, G. M., Yang, D., & Srinivasan, D. (2019). *Automatic hourly solar forecasting using machine learning models*.

Yang, D. (2020). Choice of clear-sky model in solar forecasting. *Journal of Renewable and Sustainable Energy*, *12*(2). https://doi.org/10.1063/5.0003495

Yang, D., Yagli, G. M., & Quan, H. (2018). Quality Control for Solar Irradiance Data. *International Conference on Innovative Smart Grid Technologies, ISGT Asia 2018*, 208–213. https://doi.org/10.1109/ISGT-Asia.2018.8467892

Younes, S. (2006). *Improved quality control procedures and models for solar radiation using a world-wide database*.

Younes, S., Claywell, R., & Muneer, T. (2005). Quality control of solar radiation data: Present status and proposed new approaches. *Energy*, *30*(9 SPEC. ISS.), 1533–1549. https://doi.org/10.1016/j.energy.2004.04.031

Zhang, T., Stackhouse, P. W., Gupta, S. K., Cox, S. J., Colleen Mikovitz, J., & Hinkelman, L. M. (2013). The validation of the GEWEX SRB surface shortwave flux data products using BSRN measurements: A systematic quality control, production and application approach. *Journal of Quantitative Spectroscopy and Radiative Transfer*, *122*, 127–140. https://doi.org/10.1016/j.jqsrt.2012.10.004

Declaration

I hereby confirm that I, Alexandra Reiß, have written this thesis independently and that I have not used any auxiliary materials other than those indicated. The passages of the work, which are taken from other works (including Internet sources) in terms of wording or meaning, have been marked with an indication of the source. Furthermore, I confirm that this thesis has not yet been submitted as part of another examination process neither in identical nor in similar form.

_____ _____

Place, Date Signature