WIRTSCHAFTSUNIVERSITÄT WIEN
Vienna University of Economics and Business

Bachelor´s Thesis

Title of Bachelor´s Thesis	Assessing the reliability of solar irradiance forecasts

Author (last name, first name):	Banov, Aleksandar
Student ID number:	12030634
Degree program:	Bachelor of Business and Economics, BSc (WU)
Examiner (degree, first name, last name):	Ph.D. Marta Sabou, Msc. Katrin Schreiberhuber

I hereby declare that:

1. I have written this bachelor´s thesis myself, independently and without the aid of unfair or unauthorized resources. Whenever content has been taken directly or indirectly from other sources, this has been indicated and the source referenced.
2. This bachelor´s thesis has not been previously presented as an examination paper in this or any other form in Austria or abroad.
3. This bachelor´s thesis is identical with the thesis assessed by the examiner.
4. (only applicable if the thesis was written by more than one author): this bachelor´s thesis was written together with

The individual contributions of each writer as well as the co-written passages have been indicated.

10.07.2024
Date

AleksBanov
Signature of student

Bachelor Thesis

Assessing the reliability of solar irradiance forecasts

Aleksandar Banov

Date of Birth: 19.11.2001
Student ID: 12030634

Subject Area: Information Business

Studienkennzahl: UJ 033 560

Supervisor: Marta Sabou, Katrin Schreiberhuber

Date of Submission: 10. July 2024

Department of Information Systems & Operations Management, Vienna University of Economics and Business, Welthandelsplatz 1, 1020 Vienna, Austria

Contents

1 Introduction **7**
1.1 Thesis Objective 9

2 Background and Related Work **10**
2.1 Solar irradiance fundamentals 10
2.2 Probabilistic vs deterministic forecasts 11
2.3 Evaluation of point forecasts 11
2.3.1 Measure-oriented approach 11
2.3.2 Distribution-oriented approach 13

3 Methodology of forecast quality evaluation **14**
3.1 General Approach 15
3.2 Quantitative verification based on the Murphy-Winkler framework 16
3.2.1 Bias variance decomposition 17
3.2.2 Calibration-refinement based MSE decomposition . . . 17
3.2.3 Likelihood-base rate based MSE decomposition 18

4 Implementation and Results **18**
4.1 Data 18
4.2 Stratification 20
4.3 Quantitative assessment of the joint distribution 20
4.4 Visual assessment the joint distribution 21
4.5 Marginal distribution Analysis 23
4.6 Visual analysis of the conditional distribution 25
4.7 Quantitative analysis of the conditional distribution 29
4.8 Linear variance correction 30

5 Discussion and Conclusion **33**

List of Figures

1 Solar PV Global Capacity and Annual Additions, 2012-2022 [30]. 8

2 Examples of two different methods giving the same score of RMSE on the same day but with different behavior [34]. . . . 13

3 Scatter plots for all data, temperature and wind speed bins. Identity line is in red, 2D density contours are in blue. 23

4 Marginal distributions of $p(x)$ and $p(f)$ and the Wasserstein metric. 25

5 Conditional distributions of $p(x)$ and $p(f)$ showing the variability of x given f and the variability of f given x. Black triangles mark the values at which the centroids should lie, peaks are marked by the yellow dots. 27

List of Tables

1 General error metrics for different temperature bins. 22
2 General error metrics for different wind speed bins. 22
3 Error terms based on bias-variance decomposition according to temperature. 25
4 Error terms based on bias-variance decomposition according to wind speed. 26
5 Error terms based on calibration-refinement factorization according to temperature. 28
6 Error terms based on calibration-refinement factorization according to wind speed. 28
7 Error terms based on likelihood-base rate factorization according to temperature. 30
8 Error terms based on likelihood-base rate factorization according to wind speed. 30
9 Corrected error terms based on calibration-refinement and likelihood-base rate factorization according to temperature. 32
10 Corrected error terms based on calibration-refinement and likelihood-base rate factorization according to wind speed. 33
11 Original vs. corrected general error metrics for different temperature bins. 33
12 Original vs. corrected general error metrics for different wind speed bins. 33

Abstract

In light of progressive legislative support and technological advancements, solar photovoltaic (PV) technology has emerged as a pivotal component of the global shift towards renewable energy sources and now represents a substantial portion of new renewable capacity additions worldwide. Efficient integration of solar power is challenging due to the intermittent nature of solar irradiance and calls for precise and reliable prediction models, particularly for global horizontal irradiance (GHI). This thesis investigates the reliability of deterministic solar irradiance forecasts at 24-hour forecast horizons. The measure-oriented approach and the Murphy-Winkler framework-based distribution-oriented approach are both utilized to provide a nuanced understanding of forecast quality, focusing on aspects such as calibration, resolution, and discrimination. This analysis specifically examines how forecasts perform under varying weather conditions, with the particular emphasis falling on temperature and wind speed. Additionally, a post-processing step involving a linear variance correction is implemented to refine the results.

1 Introduction

In recent decades, due to progressive laws and technological advancements, the use of renewable energy sources has significantly increased worldwide. One of the most extensively applied of these has been solar photovoltaic (PV) technology. The introduction of novel materials and notable improvements in cell efficiency have all contributed to the significant advancement of PV technology. Additionally, advances in manufacturing techniques have led to improvements in performance and decreased costs [3], [35]. The solar photovoltaic industry is growing quickly, and in 2023, solar PV alone accounted for three-quarters of the renewable capacity additions around the world. As a critical part of the transition to sustainable energy, this pattern highlights how appealing and competitive solar photovoltaics are becoming [1]. Figure 1 shows solar PV global capacity and annual additions for 2012-2022.

Solar photovoltaics offers certain advantages over conventional fossil fuel-based electricity sources. Due to economies of scale and rapid technological advancements, solar PV is a proven technology that has become remarkably cost-effective [31]. This has led to today, where more than half of new solar PV plants offer cheaper power than existing fossil fuel facilities [1]. Solar PV is anticipated to play a major role toward the Net Zero by 2050 ambition. The nature of solar PV is modular and enables varying deployments. These can range from small rooftop installations to large-scale plants. This makes the technology very flexible, and, in combination with the continuous cost reductions, solar PV is expected to be one of the key technologies in decarbonizing the global energy system [31]. Currently, there are notable gaps among various renewable technologies, and while a combination of many is required to meet the Net Zero targets, annual additions of solar PV appear to follow a more promising trajectory compared to wind, hydropower, and other renewables [1].

This surge in reliance on solar energy brings to the forefront the importance of accurate solar irradiance forecasting.The integration of solar power into the power grid presents a challenge because of its intermittent nature, necessitating reliable forecasting methods. Forecasting methods encompass a range of approaches, including data-driven approaches,image-based approaches, numerical weather prediction (NWP) models, and hybrid approaches. Each method's suitability varies based on the forecast horizon [17].

The emphasis on Global Horizontal Irradiance (GHI) (see Section 2.1 for a description of GHI) forecasting emerges as a critical aspect in solar energy research. The increasing integration of solar power into electrical systems has led to a heightened need for precise and reliable prediction models,

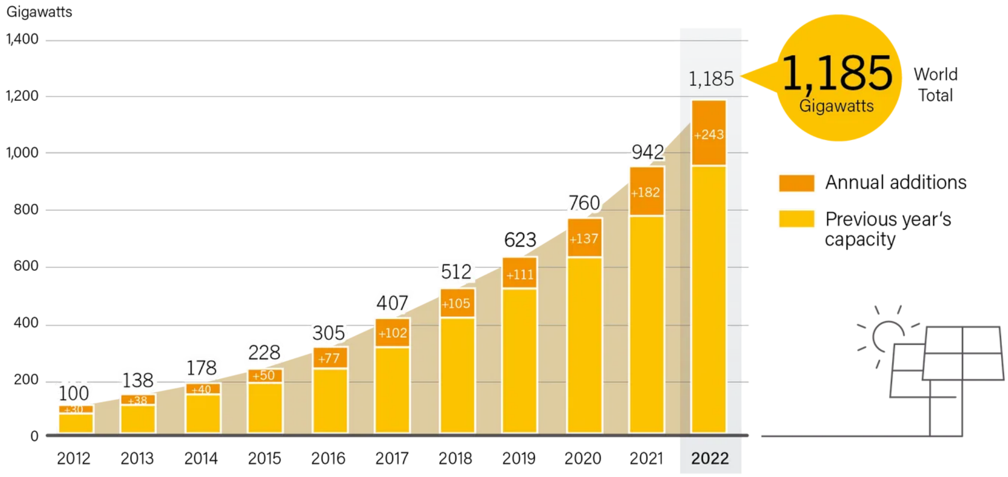


Figure 1: Solar PV Global Capacity and Annual Additions, 2012-2022 [30].

particularly for GHI. This form of irradiance forecasting has a prominent role in various PV power prediction systems [29]. There isn't much difference between GHI and other types of solar irradiance when it comes to the methodology of forecasting them. However, GHI forecasts tend to be more accurate because the variability of GHI tends to be less pronounced [38].

In the dynamic landscape of electricity markets, the precision of solar irradiance forecasts, directly influences the economic benefits for end users. The penalties for deviations from scheduled production can vary considerably. Thus, the ability to accurately predict solar power output a day in advance becomes a vital economic strategy, where more accurate forecast models can lead to substantial economic gains. However, the relationship between forecast accuracy and market conditions is more nuanced than lower error rates equating to higher economic returns, as market dynamics play a significant role in the actual economic impact of forecast precision [4].

In recent times, significant progress in the field of solar forecasting has only started to be made in the 2010s. In contrast, other areas of energy forecasting, such as load forecasting, have been utilized for planning for more than a century. Over the past decades, researchers and practitioners have begun to pay more attention to short-term load forecasting as power firms have focused on optimizing their operations. Load forecasters have had the opportunity to experiment with numerous forecasting methods over the past few decades [12]. A significant amount of effort has been devoted to the development of models for solar irradiance or solar power generation. Some of them can be found here ([13], [14], [28], [21], [18], [23], [8]). Now that solar forecasting is at the forefront, one important issue has to do with the "myth of the best technique".

Hong and Fan [11] draw attention to this problem by pointing out that although scientists have long searched for the most precise forecast, the quest for a single, "best" method is inherently faulty. They examined a wide range of methods used in load forecasting, including hybrid models and neural networks. Although some hybrid methods were useful, the majority only made a small difference. This pattern is representative of a larger trend in energy forecasting, where the emphasis frequently switches to integrating different approaches in the pursuit of developing better hybrid solutions. Unfortunately, this often results in models that are hard to duplicate and generalize, which hinders the development of useful forecasting applications [11].

Forecast verification relates to the quality of a forecast. For a given forecast, there are numerous different numerical scores that can be used to calculate the relative forecast quality. Examples of such error metrics are mean bias error (MBE), mean absolute error (MAE), and root mean square error (RMSE) [10]. This measure-oriented approach has been widely used in solar forecasting. An alternative to this measure-oriented approach was proposed decades ago, where the joint distribution of forecasts and observations can be used to examine the skillfulness of the forecasts. This distribution-oriented approach was proposed by Murphy and Winkler [27]. It gained significant popularity in the field of weather forecasting but has only very recently been applied to the domain of solar irradiance forecasting, first with the work of Yang and Perez [41] and subsequent works [20], [38], [39], [42]. The distribution-oriented approach goes beyond the fundamental aspects of accuracy and skill and presents the problem as multi-dimensional. Factoring the joint distribution into a marginal and a conditional distribution allows us to explore more aspects of forecast quality like reliability, resolution, and discrimination [25].

1.1 Thesis Objective

The objective of the thesis is to evaluate the quality of deterministic global horizontal irradiance (GHI) forecasts at 24-hour horizons by comparing them to actual weather conditions for a specific geographical location. Furthermore, the analysis includes how the quality of the forecasts is influenced by climatic effects or weather variables such as temperature and wind speed. As mentioned above, the quality of the forecast is expressed in different aspects, like reliability, resolution, and discrimination.

2 Background and Related Work

The forthcoming sections will provide an overview of fundamental concepts around solar irradiance and the methods used for the evaluation of solar forecasts. Section 2.1 introduces the core concepts of solar irradiance, including its primary components and their roles in solar energy systems. Following this, Section 2.2 contrasts the methodologies and implications of deterministic forecasts with the broader scope of probabilistic approaches, highlighting the need to account for uncertainties in forecasts. Finally, Section 2.3 outlines the assessment of solar irradiance forecasting accuracy through traditional metrics like RMSE, MAE, and MBE, focusing on some of their strengths and limitations in capturing forecast quality. It further expands into a short introduction of the distribution-oriented approach following the Murphy-Winkler framework, which evaluates forecasts beyond basic error measures and offers a more comprehensive perspective on the effectiveness of forecasts.

2.1 Solar irradiance fundamentals

Extraterrestrial radiation (ETR) is the total electromagnetic radiation that is emitted from the sun [2]. As it travels towards the earth's surface, it decreases due to absorption, reflections, and re-emissions caused in the atmosphere and is broken down into two components: diffuse horizontal irradiance (DHI) and direct normal irradiance (DNI). DNI is the amount of solar radiation that comes straight from the sun and strikes a surface perpendicular to the sun's rays. DHI, or diffuse horizontal irradiance, refers to the solar radiation that has been scattered by the atmosphere and reaches the surface from all directions. [22], [17]. Global horizontal irradiance (GHI) is the geometric sum of these two components and represents the combined total of all solar radiation striking a horizontal surface:

$$\mathrm{GHI} = \mathrm{DHI} + \mathrm{DNI} \cdot \cos(\theta), \tag{1}$$

where θ is the solar zenith angle.

GHI is applicable in PV systems, and DNI is applicable in concentrated solar power plants (CSPs) [15]. This is because most CSPs are only able to effectively concentrate DNI, while photovoltaic systems can make use of both DNI and diffuse horizontal irradiance [22]. Cloud coverage is the main reason for the reduction in both types of solar irradiance. However, when clouds are not present, aerosol concentration becomes the major factor in the reduction of intensity. The reduction in DNI intensity can vary from 30% to 100% based on the exact aerosol concentration, while for GHI the reduction is considerably lower at about 10% [19], [22]. This causes forecasts of DNI to be notably less accurate than GHI, even when methodologically there isn't any significant difference between the prediction of the two [38].

2.2 Probabilistic vs deterministic forecasts

Conventional solar power forecasting generates a single value or the conditional expectation of solar power output at a future time point, known as a "deterministic" or "point forecast." However, predictions involve uncertainty, which probabilistic forecasting can address, considering time and space dimensions [9]. Probabilistic forecasts can be presented as probability distributions, quantiles, or intervals, while point forecasts, or single-valued forecasts, offer summarized statistics, mainly expected values over different periods. In weather forecasting, it is well known that forecasts span three-dimensional space, time, and probability [12]. Probabilistic forecasts provide valuable information on forecast uncertainty, essential for communicating events with potential significant losses. However, understanding probabilistic forecasts can be more challenging than single-value point forecasts. Some users prefer clear, definitive statements rather than making optimal decisions themselves. In reality, this can lead to probabilistic forecasts being used to shift the responsibility for decision-making from the forecasting community to the user community [7].

2.3 Evaluation of point forecasts

Here we introduce the measure-oriented approach for verification of solar irradiance forecasts using metrics like RMSE, MAE, and MBE to assess accuracy and bias. We give some overview of the implications of these metrics in representing error distributions and their limitations in fully capturing the complexity of forecast accuracy. And transition into a brief description of the distribution-oriented approach, which utilizes the joint distribution of forecasts and observations to provide a deeper analysis of forecast skillfulness, moving beyond traditional error metrics to consider the forecasts reliability, resolution, and discrimination.

2.3.1 Measure-oriented approach

Verifying irradiance forecasts is much like verifying other meteorological variables, with slight differences that have to be taken into account. After generating a solar irradiance forecast using a specific method, its effectiveness can be evaluated using a validation dataset comprising historical forecasts and observations. For point forecasts, the quality is commonly validated in the literature using error metrics such as root mean square error (RMSE), mean absolute error (MAE), and mean bias error (MBE) [9]. The RMSE provides a global error measure throughout the entire forecasting period. It is given by:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{I}_i - I_i)^2}, \tag{2}$$

where $\hat{I}_i$ is the predicted value, I_i is the observed value, and N is the total number of observations. Because each error term is squared, the RMSE metric effectively weights large errors more heavily than small errors, which tends to penalize large forecast errors. Similar to the RMSE metric, the MAE metric also measures global errors, but it does not penalize extreme forecast events as severely. Better forecasts are indicated by smaller MAE values. A limitation of the MAE metric is that a small number of large errors can be easily overwhelmed by a large number of very small errors. This may pose an issue for systems that are susceptible to extreme weather events. It is given by:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{I}_i - I_i\right|, \tag{3}$$

where $\hat{I}_i$ is the predicted value, I_i is the observed value, and N is the total number of observations. The MBE metric aims to show the average forecast bias. The mean bias error (MBE) is given by:

$$\text{MBE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{I}_i - I_i), \tag{4}$$

where $\hat{I}_i$ is the predicted value, I_i is the observed value, and N is the total number of observations [44], [17], [10]. The larger the MBE, the larger the forecasting bias. A positive MBE indicates over-forecasting, whereas a negative MBE indicates under-forecasting, assuming that the forecast error is equal to the forecast minus the actual power generation. The entire range of forecast errors is not well indicated by MBE. For instance, multiple significantly distinct error distributions, some of which might be more advantageous than others, could be represented by the same MBE value. This is, however, not a limitation of MBE specifically. Most forecasting metrics are unbiased only if the underlying error distribution is Gaussian [44].

It has been established that no single metric stands paramount when it comes to assessing the quality of a forecast. A whole array of metrics can

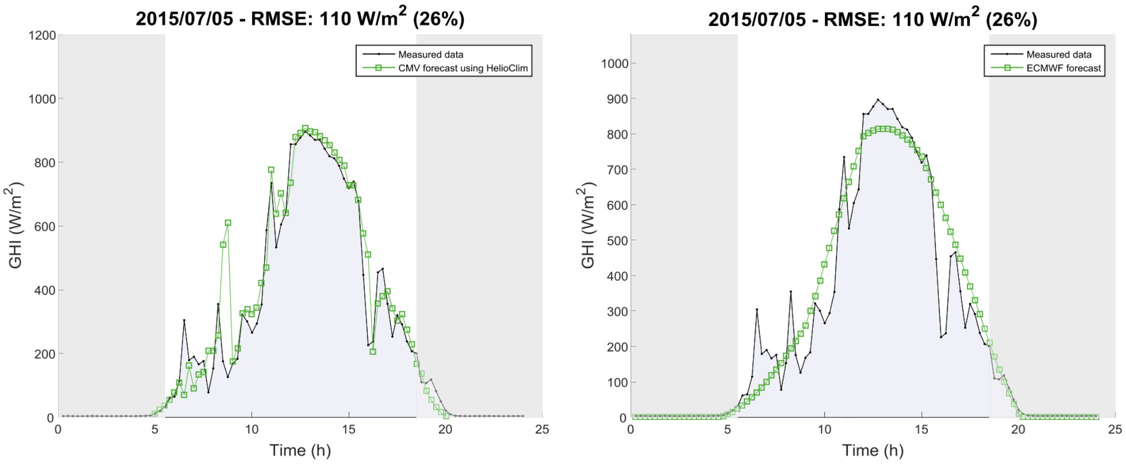


Figure 2: Examples of two different methods giving the same score of RMSE on the same day but with different behavior [34].

be used together to verify various aspects in a more intricate manner. While using a comprehensive range of metrics might lead to some of them having overlapping elements in terms of what they measure, they all can be a part of a suite that collectively offers a robust toolkit for the assessment of solar forecasting [44]. Nevertheless, these metrics have the limitation of underdetermination. Even when many of these metrics are used in combination, it is simple to confirm that they do not describe unique error characteristics. In actuality, these metrics can yield identical values for a wide variety of combinations of error characteristics [33]. This means that forecasts with exactly the same error terms can have vastly different distributions. Figure 2 serves as an illustration of how two sets of quite disparate forecasts can have the same error. These measures omit additional information about the forecast errors that may be of significant importance since they are unable to distinguish between two distributions with the same mean and variance but differing skewness and kurtosis values [44].

2.3.2 Distribution-oriented approach

An alternative to this measure-oriented approach has been proposed decades ago by Murphy and Winkler [27], where the joint distribution of forecast and observation can be used to examine the skillfulness of the forecasts. Determining the overall quantity of information a forecaster has access to during verification is helpful, as forecast quality analysis is influenced by the data contained in forecast-observation pairings. A forecaster is no longer constrained by the collection of summary statistics by defining all available data. Put another way, the joint distribution of forecast and observation can be used to examine the forecasts' skillfulness if the temporal sequence of forecast-observation pairings is not of concern because the data included is time-independent [38]. The use of the join distribution to evaluate the quality of different forecasts has been popular, especially in meteorology [41]. Among the different methods to describe the joint distribution, Murphy and Winkler's forecast-verification framework is arguably the most general, as it applies to both deterministic and probabilistic forecasts [27]. It has gained significant popularity in the field of weather forecasting but has only very recently been applied to the domain of solar forecasting.

This framework facilitates easier access to the information within the joint distribution by decomposing the mean square error (MSE), which is essentially the squared RMSE [41]. Enhanced by Bayes' theorem, the Murphy-Winkler framework represents the joint distribution as the product of marginal and conditional distributions, making the embedded information more accessible [38]. These decomposed metrics evaluate observation-forecast pairs from various aspects that define good forecasts, such as reliability, resolution, and discrimination, which can be easily defined, quantified, and most importantly, interpreted in a systematic way. For a more detailed overview of those aspects and the Murphy-Winkler framework in general, see Sections 3.1 and 3.2.

3 Methodology of forecast quality evaluation

This section introduces the general problem of the evaluation of forecast data and outlines our method, which combines measure-oriented and distribution-oriented approaches to comprehensively analyze forecast quality. The section underscores the synergy between visual and quantitative analysis, emphasizing how distribution-oriented insights are not an alternative but an enhancement to the traditional metrics-based approach. In Section 3.2, the focus shifts to detailing how MSE can be decomposed to further elucidate distinct aspects of forecast accuracy based on the bias-variance decomposition, the calibration-refinement, and likelihood-base rate factorizations. This detailed examination enables a deeper understanding of how the Murphy-Winkler framework works to describe forecast performance under various conditions, enhancing the overall assessment process.

3.1 General Approach

The evaluation of the forecast data will feature both the measure-oriented and the distribution-oriented approach. This is because these two approaches are, in actuality, complementary to each other, and one is never enough to completely substitute the other. In fact, at the start of the analysis, we include a forecast-observation scatter plot to check forecast quality. We note the distribution of the cloud points in relation to the identity line. In effect, what we observed is exactly the joint distribution of forecasts and observations. Subsequently, when error metrics like MBE, RMSE, NRMSE, and MSE are calculated, they also represent the joint distribution [38]. Meaning that these measures are just summaries of the joint distribution. Consequently, when the measure-oriented approach employs a visual accuracy quantification or one based on accuracy measures, the joint distribution is always present; however, even more insight can be drawn from the distribution-oriented approach.

If the forecasts are f and the observations x, the joint distribution is denoted by $p(f, x)$ and consists of relative frequency occurrence of specific combinations of forecast and observation values. Although the joint distribution of forecasts and observations contains all information relevant to verification, the information is more accessible when the distribution is factored. Following the Bayes rule, any joint distribution can be factored into a conditional and a marginal distribution in two ways. Thus, we obtain two factorizations that reveal information that has been embedded into the joint distribution and relates to particular aspects of verification. The first factorization is denoted by the conditional distributions of the observations given the forecasts $p(x|f)$ and the marginal distribution of the forecasts $p(f)$:

$$p(f, x) = p(x|f)p(f). \tag{5}$$

The conditional distribution $p(x|f)$ indicates how often different observations have occurred when a particular forecast was given, or in other words, the variability of the observations given a particular forecast. If a specific value is forecasted multiple times by a model, there is no rationale for expecting the materialized observation to always be the same, as that would signify that the forecast is completely deterministic. In reality, that is not the case, and all the observations given a forecast value form a distribution $p(x|f)$. If, given a particular forecast value, the mean of those observations is equal to the forecasted value, then the forecasting model would be considered perfectly calibrated. Or, in mathematical terms, $E(x|f) = f$. The marginal distribution $p(f)$ indicates how often different forecast values are used. If the same forecast value is always predicted, the forecasting model is said to not be refined or sharp. Here, we see that both calibration and refinement are distinct concepts that are both of interest for verification purposes. A forecast can be perfectly calibrated but also have no refinement. Such a forecast will always predict the mean of the observations. Complete refinement is difficult to define for deterministic forecasts; however, $p(f)$ has to equal $p(x)$ if we were to have a perfect forecast [26].

The second factorization involves the conditional distributions of the forecasts given the observations $p(f|x)$ and the marginal distributions of the observations $p(x)$:

$$p(f, x) = p(f|x)p(x). \tag{6}$$

The conditional distribution $p(f|x)$ indicates how often different forecast values are produced before a specific value of x is observed, or, in other words, the variability of the forecasts, given a particular observation. This aspect is known as the likelihood, and $p(f|x)$ indicates how well the forecast discriminates between different values of observations. If $p(f|x)$ is zero for all values x but one, the forecast is perfectly discriminatory. If $p(f|x)$ is the same for all values of x, the forecast is not at all discriminatory. In reality, we would like to see predictions that have a high concentration around the specific x. The best forecast provides us with likelihoods that are very different for different observations, which would mean that it is very discriminatory and informative about x. The marginal distribution $p(x)$ indicates how often different values of x occur. It is known as the base rate. Since it is the only element that does not involve f in any way, it is descriptive not of the forecasting model but of the forecasting situation. Analyzing the distribution of observations gives insight into the situation that the forecasting model is trying to predict. If the distribution is peaked, there is relatively little uncertainty, and forecasting is easier. If the distribution is uniform, the uncertainty is high.

Given that the two factorizations' components clearly measure distinct aspects of the forecasting system and/or forecasting circumstances, each of the four components will be very valuable for verification.

3.2 Quantitative verification based on the Murphy-Winkler framework

Because many of the error metrics, like MBE, MAE, and RMSE, are just different ways of summarizing the joint distribution, an error metric like MSE can be decomposed multiple ways into various terms, with each term describing a distinct aspect of forecast quality. Three decompositions will be considered: the bias-variance decomposition and decompositions based on the calibration-refinement and likelihood-base rate factorizations. Their derivations are found in the appendix of [24]. The latter two decompositions are directly connected to the Murphy-Winkler framework, while the first has been widely popular in a broader statistical context for decades. The sections below provide details on all three; for further information, they are based on the following references: [27], [41], [26], [39], [42], [38].

3.2.1 Bias variance decomposition

The mean squared error (MSE) can be expressed as:

$$\mathrm{MSE} = V(f) + V(x) - 2\operatorname{cov}(f, x) + [E(f) - E(x)]^2. \tag{7}$$

Here $V(f)$ and $V(x)$ describe the variance of the predictions and observations and are summaries of the marginal distributions. The covariance is a measure of the linear relationship between the forecasts and observations, which is termed the association. The last term $[E(f) - E(x)]^2$ is the squared unconditional bias, or MBE^2. None of the terms here feature the conditional distribution of forecasts and observations on their own, so for an interpretation of it, we turn to the calibration-refinement and likelihood-base rate MSE decompositions.

3.2.2 Calibration-refinement based MSE decomposition

The mean squared error (MSE) can additionally be expressed as:

$$\mathrm{MSE} = V(x) + \mathrm{E}_f\left[f - \mathrm{E}(x|f)\right]^2 - \mathrm{E}_f\left[\mathrm{E}(x|f) - \mathrm{E}(x)\right]^2. \tag{8}$$

Here E_f denotes the expectation with respect to the marginal distribution $p(f)$, and $E(x|f)$ is the conditional expectation of x on f. $E_f[f - E(x|f)]^2$ can be calculated by taking the mean of $[f - E(x|f)]^2$ which is obtained by evaluating the mean of the conditional distribution $E(x|f)$ for each given value of f. The other terms of both decompositions are calculated in a similar manner.

In Eq. (8), the first term is the variance of the observations, which is a quantification of the marginal distribution of observations, also known as the base rate or uncertainty. The second term $E_f[f - E(x|f)]^2$ relates to calibration or reliability. It describes the degree of correspondence between the mean observation given a particular forecast and the forecast associated with that observation. The third term $E_f[E(x|f) - E(x)]^2$ relates to the difference between the conditional expectations and the marginal expectation. This is known as the resolution of a forecast, and the negative sign means that larger values are preferred. This is because if $E(x|f) = E(x)$ is being approached, different forecasts would be followed by very similar observations, and the forecast would have very little meaning.

3.2.3 Likelihood-base rate based MSE decomposition

The mean squared error (MSE) can be also expressed as:

$$\mathrm{MSE} = V(f) + \mathrm{E}_x\left[x - \mathrm{E}(f|x)\right]^2 - \mathrm{E}_x\left[\mathrm{E}(f|x) - \mathrm{E}(f)\right]^2. \qquad (9)$$

In Eq. (9), the term $E_x[(x - E(f|x)]^2$ relates to the degree of correspondence between the mean forecast given a particular observation and the observation associated with that forecast. It is known as type 2 conditional bias and should be minimized. It can be viewed as the weighted average of the errors in the average forecast. The last term $E_x[(E(f|x) - E(f)]^2$ is the weighted square difference between the average forecast associated with each observation and the overall average of forecasts. Larger differences are preferred, as the term measures the extent to which an average forecast associated with an event differs from the average forecast; this is known as discrimination and needs to be maximized.

4 Implementation and Results

In Section 4, we address the details around the implementation of the forecast analysis, beginning with data collection, spatial and temporal matching of data, followed by stratification based on the weather variables of temperature and wind speed. Quantitative assessments using traditional error metrics (MBE, RMSE, and MAE) evaluate the joint distribution, while visual assessments expand that to examine both the marginal and the conditional distributions of forecasts and observations. Quantitative analysis based on the Murphy-Winkler factorizations follows to outline the more nuanced aspects of forecast quality like calibration, type 2 conditional bias, resolution, and discrimination. A post-processing step in the form of a linear variance correction scheme is implemented to address disparities in variance present within the stratified data, providing a more balanced comparison across different weather states.

4.1 Data

Regardless of the purpose behind forecast verification, it always starts with a matched set of observations and forecasts. This occurs both spatially and temporally. Two approaches are relevant in the context of spatial matching, namely gridded observations and station observations. While they both have certain advantages and disadvantages, their use mainly comes down to the type of verification. Model-oriented verification works best with gridded observations, while the specificity of point observations is relevant when the accuracy for a particular location is of interest [6]. Our analysis focuses on the evaluation of point forecasts for a single location.

The forecast data consists of global horizontal irradiance (GHI) values at 15-minute resolution for several days in the future. The historical observation data is from GeoSphere Austria, where the Messstationen Zehnminutendaten v2 (ZEHNMIN) [5] dataset contains station data at 10-minute resolution, with the majority of the measurement data being quality tested and accompanied by quality flags.

When temporally matching solar observation and forecast data, it is important to keep in mind the bell-shaped diurnal cycle of solar irradiance since slight time series misalignment can discredit the validity of the verification [42]. In solar forecasting, to achieve proper alignment, we often need to transform a high resolution time series into a lower temporal resolution. This is accomplished by time-oriented aggregation following one of three schemes, namely, floor, ceiling, and round. Establishing the details around which of these methods was used for a given observation dataset is important since any mismatch will exaggerate the errors during verification [37].

In our case, for the temporal alignment of the observation and forecast data, the 10- and 15-minute records will be aggregated and time stamped to the nearest half an hour. Meaning that the timestamps that make up half an hour will be averaged and aggregation will be done using the "ceiling" operator, e.g., for the station data, 10-min data points between 11:00 and 11:30 are stamped with 11:30 after aggregation. We chose the ceiling method since it matches how the 10-min resolution GHI data would be recorded in a station where the timestamp of 11:00 is based on tracking GHI sensor data between 10:50 and 11:00 and uses the latest time stamp. However, we do not find documentation for how the GHI forecasts are exactly produced and thus choose to maintain the same approach as with the observations. This presents a potential limitation within this process. If we assume the inverse, namely that the forecasts are aggregated using the floor method, a data point labeled as 11:30 would cover data collected from 11:30 to 12:00. This temporal alignment inherently shifts the bias, predisposing the forecasts to overestimate the GHI when compared with the expected observation. Special matching will be based on which station from the ZEHNMIN dataset is closest to the location of the forecast.

It is important to note that the irradiance data records contain many values at or close to 0, specifically during the early morning and late afternoon. This is one of the reasons why not every accuracy metric can be used to evaluate the quality of solar forecasts [43]. Furthermore, there is no consensus on whether or not to include overnight hours in the validation process [38]. Since the forecasts are perfectly accurate at night, including those hours would reduce the overall error. As a result, in our implementation, the data collected at night is filtered out. A zenith-angle filter of less than 85 degrees is used to guarantee this. Additionally, this provision eliminates low-sun conditions during the early morning and late afternoon, which lead to inaccurate measurements and are often of insufficient irradiance for the purpose of solar PV systems.

4.2 Stratification

After appropriately matched sets of observations and forecasts, stratification can take place. It has the purpose of dividing the samples into relatively homogeneous subsets and is key to answering specific questions regarding forecast behavior. It is common to stratify based on lead time (12h, 24h, 48h, etc.), season, geographic area, as well as other dimensions relevant to the specific parameter that is being verified [32]. After the decisions about stratification have been made, the visual and quantitative evaluation of the forecast can start.To fulfill the goal of the thesis regarding discernible patterns of forecast quality based on weather variables, the sample data will be stratified based on temperature and wind speed. These are features of the forecast dataset. An alternative is to do the stratification based on the observation, meaning defining categories according to observation values. For our implementation, the data will be split into four equally sized bins, from lowest to highest temperature and wind speed.

4.3 Quantitative assessment of the joint distribution

Mean bias error (MBE), root mean square error (RMSE), and mean absolute error (MEA) are widely used metrics not only in the solar forecasting community [41], [20], [38], [42], but also in meteorology at large [39], [16], [36]. In our case, those metrics are used as a specific way of summarizing the joint distribution. Table 1 and Table 2 tabulate those metrics as well as their normalized versions, nMBE, nRMSE, and nMAE. The normalized versions are computed by dividing the metric with the mean of observation, e.g., $\mathrm{nMBE}(f, x) = \frac{\mathrm{MBE}(f,x)}{E(x)}$. This is done because most of the summary measures are scale-dependent, meaning that if the underlying data samples differ in scale, it is very hard to interpret the values of the error metrics.

Normalization of error measures in the field of meteorology is rarely used; however, in the subfield of solar forecasting, it is much more common. The specific method of normalization by division by the mean of observations is seen in the field of solar engineering [40].

Examining nRMSE and nMAE, we note a consistent decrease in the magnitude of forecast error as we go towards higher temperature buckets. This relation gets inverted when we look at the stratification by wind speed, where, as we go towards higher and higher wind speeds, nRMSE and nMAE increase. It is important to note that if we only looked at the non-normalized versions of the metrics, this insight would be obscured for the temperature buckets. However, for the wind speed buckets, the scale-dependent measures provide us with generally the same insight. This might be the case due to wind speed not strongly affecting the scale of the GHI, even when low wind conditions tend to be better than high wind conditions for the accuracy of the forecast.

Additionally, we note that the relative decrease in nRMSE compared to nMAE is greater when comparing the highest temperature bin to the others and when comparing the lowest wind speed bin to the others. This suggests that the relative frequency and/or size of large errors are the smallest in these bins.

The faster decrease in nRMSE indicates that there are relatively larger errors in the lower temperature bins and in the higher wind speed bins. As temperature increases or wind speed decreases, these larger errors become less frequent or less significant, leading to a more pronounced reduction in nRMSE. While both nRMSE and nMAE are decreasing, indicating an overall improvement in prediction accuracy, the fact that nRMSE decreases faster implies that larger errors, which nRMSE is more sensitive to, are being reduced more significantly.

When looking at the MBE and nMBE, we see that, in general, there is a slight bias towards overestimation. However, when looking at the different values of MBE for the most accurate buckets based on the other metrics, we see that the highest temperature bucket and the lowest wind speed bucket are actually biased towards under-prediction. Further examination via scatter plots might clarify this more.

4.4 Visual assessment the joint distribution

The scatterplot, as the fundamental instrument for exploratory examination of paired data, is used here in Figure 3 to graphically represent the joint distribution itself. Where the relative frequency of (f, x) pairs is qualitatively expressed by the point density in a scatterplot of f versus x. Using 2D kernel density contours is advantageous for a large number of points.

Temp	**MBE**	**nMBE**	**RMSE**	**nRMSE**	**MAE**	**nMAE**
All	3.043418	1.051219	93.448303	32.277740	63.136697	27.355712
(-7.47 - 8.89)	2.962576	1.793277	67.096348	40.614077	48.831114	34.699949
(8.89 - 14.83)	6.674911	2.854373	87.632405	37.474001	62.467249	33.727853
(14.84 - 20.8)	4.864266	1.575032	106.190996	34.384271	71.674116	30.202132
(20.8 - 35.08)	-2.328031	-0.517066	107.121563	23.792153	69.583012	28.917183

Table 1: General error metrics for different temperature bins.

Wind	**MBE**	**nMBE**	**RMSE**	**nRMSE**	**MAE**	**nMAE**
All	3.043418	1.051219	93.448303	32.277740	63.136697	27.355712
(0.42 - 2.0)	-8.782939	-3.007534	73.386101	25.129541	52.096576	23.264058
(2.01 - 3.5)	5.875879	1.958722	98.719316	32.908040	65.748681	27.530056
(3.5 - 5.33)	9.334237	3.188756	100.988260	34.499547	66.327929	28.592532
(5.33 - 11.42)	5.753688	2.105180	97.993499	35.854206	68.380319	30.077139

Table 2: General error metrics for different wind speed bins.

When examining the most dense 2D contours in the majority of the plots, most data points lie below the identity line, except for very low values of GHI (<100). This generally indicates an underestimation. However, the positive MBE for all bins (except the highest temperature and lowest wind speed bins) would suggest overestimation, meaning that most data points should lie above the identity line. A notable pattern emerges where, for the early GHI range (0-350), we see a higher variability of errors towards overestimation. This is represented by the more spread-out 2D kernel density contours above the identity line compared to those below it. Furthermore, the points outside the 2D contours, which can be classified as outliers, are more frequent and larger above the identity line. Despite the majority of points being clustered below the identity line, these outliers contribute to the bias toward overestimation. Combined with the bias from the higher variability in lower GHI values (0-350), this causes the MBE to report overestimation even when the most dense contours of the scatter plots are almost always consistently below the identity line.

The exceptions were again the plots with the highest temperature and lowest wind speed, where MBE indicated underestimation in unison with the majority of points being below the identity line. It is also interesting to note that the scatter plot of the highest temperature bin was the only one without the pattern of higher variability of errors above the identity line for the early GHI range (0-350). Even then, the outliers were generally biased toward overestimation. For the scatter plot representing the bin of

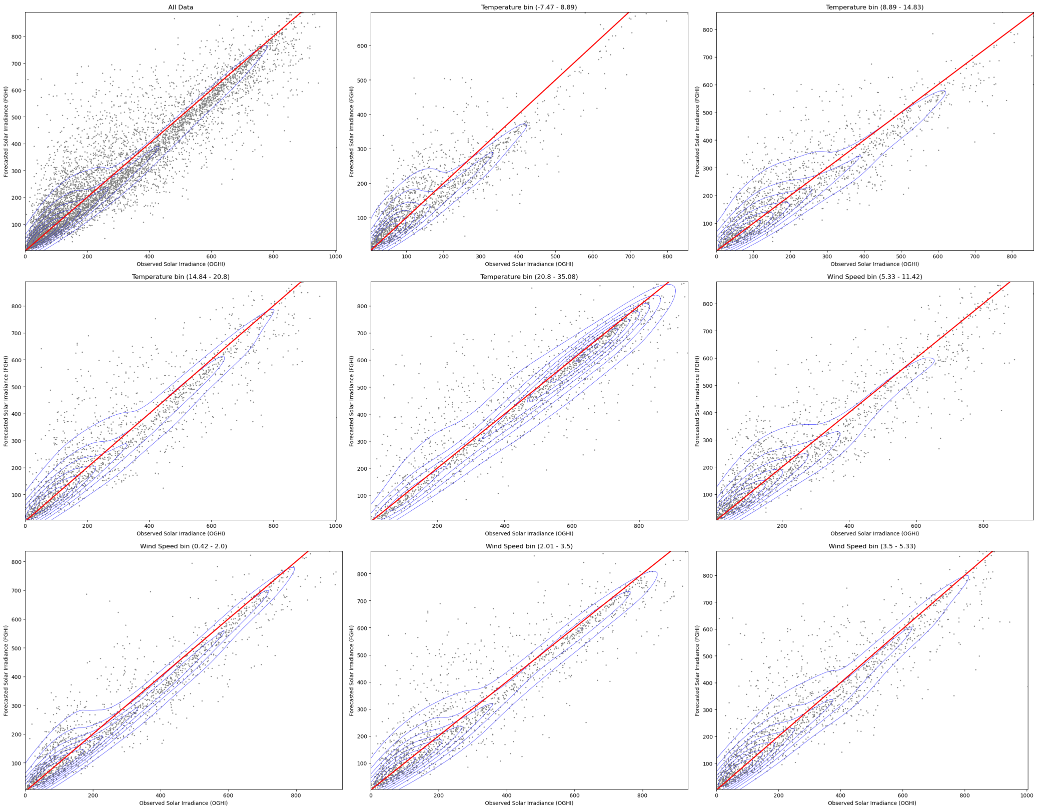


Figure 3: Scatter plots for all data, temperature and wind speed bins. Identity line is in red, 2D density contours are in blue.

lowest wind speed, we can see the same pattern for the early GHI range (0-350); however, the overwhelming majority of points below the identity line prevented the MBE from reversing its sign.

4.5 Marginal distribution Analysis

As mentioned before, if a forecast was to be perfect, the marginal distributions of $p(x)$ and $p(f)$ would be exactly the same. However, the inverse does not hold true. Meaning that if a data sample of forecasts and observations has exactly matching marginal distributions, it is not implied that the forecast is perfect and only that the distribution of predictions is consistent with observed climatology. Based on this, the analysis is most meaningful in the direction of a mismatch between the two distributions, since this is an obvious sign of a forecast that could be better.

Figure 4 represents the estimates of the continuous marginal distributions of forecasts and observations. This is why we have some plot area below zero when there are no observations or forecasts for negative GHI values. Here, the WD on the top right is the Wasserstein metric, also known as the Earth Mover's Distance. It measures the distance between two probability distributions, or more accurately, the effort required to transform one distribution into the other. Thus, the lower its value, the more the two distributions are alike.

Examining the plots, we generally see that there seems to be a slight underpopulation for the lower GHI values (<200) and a slight overpopulation for the mid-range. However, all in all, there isn't any bin with overly sharp marginal distributions, and it appears the forecast gives appropriately different forecast values with no real shrinkage in range. Interestingly, the bin with the highest temperature that would be considered most accurate appears to have the most dissimilar marginal distributions of $p(x)$ and $p(f)$. It appears that the first peak with low GHI values is underpopulated, while the peak for high irradiance between around 500 and 700 GHI is overpopulated.

Complementing this visual analysis, we quantitatively verify the marginal distributions using the widely recognized bias-variance decomposition of the mean squared error (MSE):

$$\mathrm{MSE} = V(f) + V(x) - 2\operatorname{cov}(f, x) + [E(f) - E(x)]^2. \tag{10}$$

Based on the above, it's clear that MBE can be expressed in terms of the means and variances of the marginal distributions, $p(x)$ and $p(f)$, and the covariance of the joint distribution, $p(f, x)$. The covariance can also be written as $\sqrt{V(f)V(x)}\rho(f, x)$, where $\rho(f, x)$ is the correlation between f and x. Correlation measures the linear relationship between two variables, i.e., here it measures the association of observations with forecasts. The calculation of each term is tabulated in Table 3 for temperature and Table 4 for wind speed.

Examining the variances, we see that $V(x)$ is larger than $V(f)$ for all bins of both temperature and wind speed, which indicates slight under-dispersion, where the variability of the irradiance is not fully captured by the forecasts. The linear relationship between f and x reveals that for situations with little to no wind, the association of observations with forecasts is stronger than for situations with more wind; however, further increases in wind have much less of an impact. For temperature, the correlation is highest in high-temperature circumstances, with states of lower temperature slightly decreasing the association. It is worth noting that the bias term $[E(f) - E(x)])$ has minimal contribution to the overall MSE value and is vastly overshadowed by the variances of f and x. This is quite common in state-of-the-art operational solar forecasts, and a minimized bias term is to be expected since many models

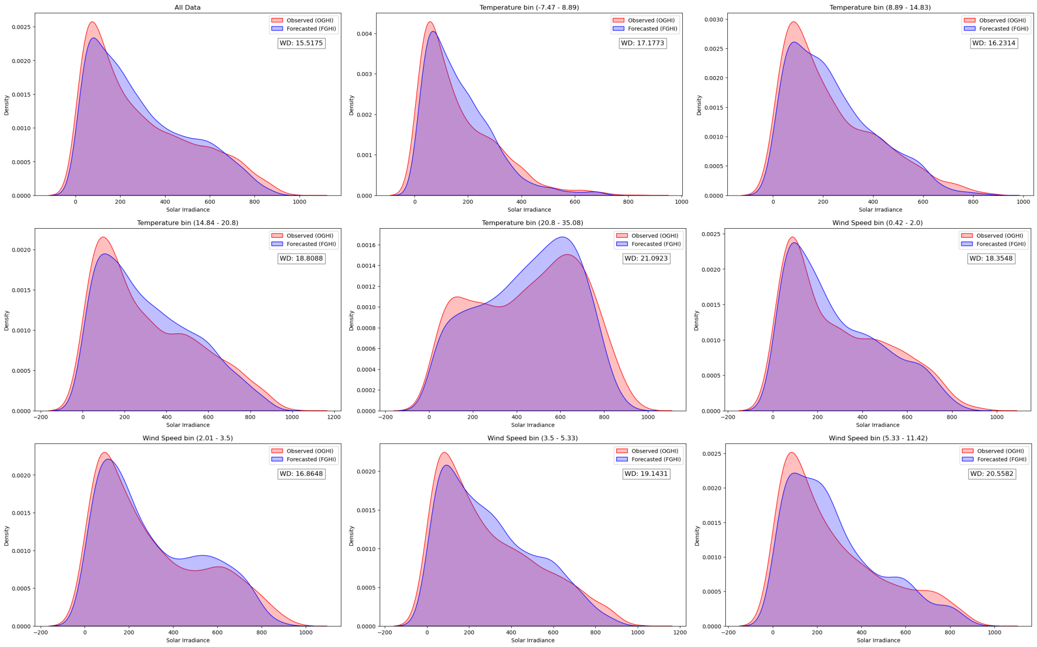


Figure 4: Marginal distributions of $p(x)$ and $p(f)$ and the Wasserstein metric.

implement a bias correction, e.g., model output statistics (MOS) [38]. This correction transforms the third term, which is essentially the MBE squared, into more of a baseline requirement than a metric for quality evaluation.

Temp	**MSE**	$V(x)$	$V(f)$	**Correlation**	**Bias**^{2}
All	8732.585	53268.157	46058.156	0.9146	9.262
(-7.47 - 8.89)	4501.920	19803.210	15310.078	0.8793	8.777
(8.89 - 14.83)	7679.438	34302.627	29240.461	0.8827	44.554
(14.84 - 20.8)	11276.528	56318.291	48533.337	0.8951	23.661
(20.8 - 35.08)	11475.029	57902.121	48395.830	0.8957	5.420

Table 3: Error terms based on bias-variance decomposition according to temperature.

4.6 Visual analysis of the conditional distribution

Having analyzed the joint and marginal distributions of forecasts and observations, we turn to the conditional distribution for further insights. Generally, the information within the conditional distributions is not part of traditional forecast verification; however, it contains several distinguishing char-

Wind	**MSE**	$V(x)$	$V(f)$	**Correlation**	**Bias**^{2}
All	8732.585	53268.157	46058.156	0.9146	9.262
(0.42 - 2.0)	5385.520	50147.283	42847.004	0.9458	77.140
(2.01 - 3.5)	9745.503	57037.423	50735.240	0.9115	34.526
(3.5 - 5.33)	10198.629	53813.055	46722.880	0.9017	87.128
(5.33 - 11.42)	9602.726	51687.939	43393.441	0.9028	33.105

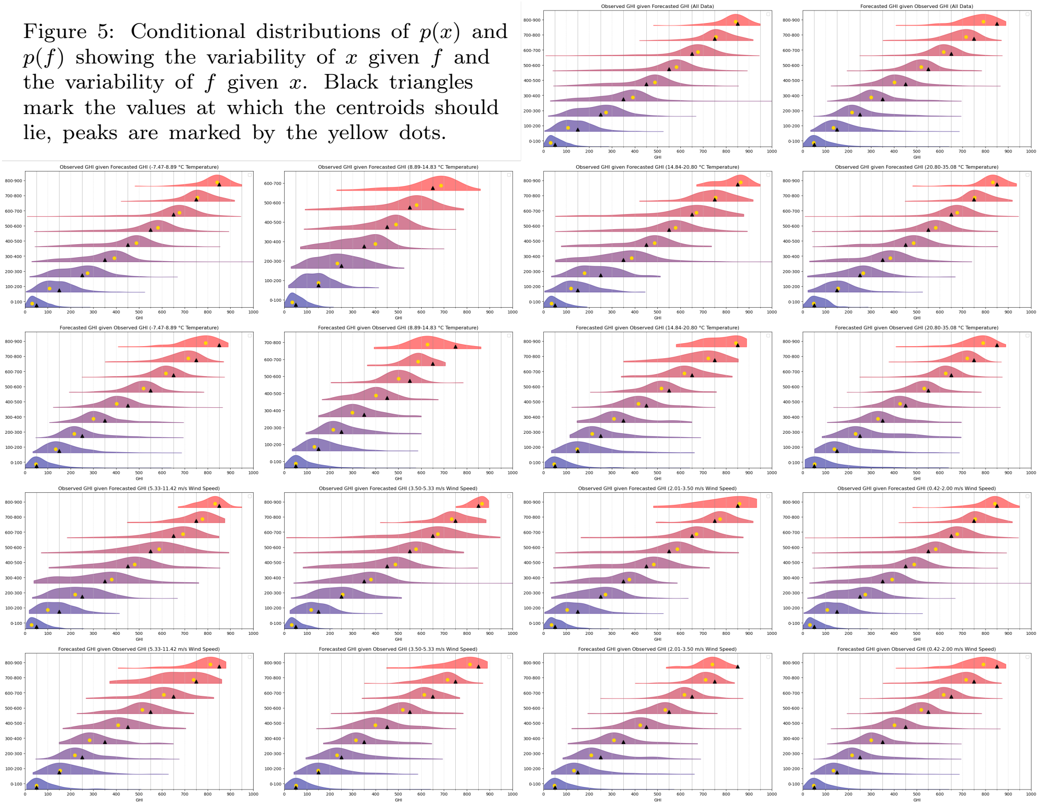
Table 4: Error terms based on bias-variance decomposition according to wind speed.

acteristics that relate to forecast quality. It is important to note that $E(x|f)$ represents the set of conditional distributions of the observations given the forecast, and in turn, $E(f|x)$ represents the set of conditional distributions of the forecasts given the observation. Because we are dealing with sets of distributions, the visual analysis of the conditional distribution of forecasts and observations is not as straightforward as the marginal distribution.

Here, rideline plots are used to show the variability of observations given a forecast and the variability of the forecast given observations for the whole range of GHI. Every plot has multiple rows; each row represents 100 units of GHI, and a kernel density estimation is done with a Gaussian kernel. Note that variability is shown only in relation to the x-axis. Normally, we have the x-axis for observations and the y-axis for forecasts. For the Figure 5 plots, that changes in regards to what is "given". For example, when we have plotted "Observed GHI given forecasted GHI", since variability can be expressed only on the x-axis, the y-axis is forecast GHI and the x-axis is observations. Conversely, when plotting "Forecasted GHI given observed GHI", the roles of the axes are reversed: the x-axis represents the forecasts while the y-axis captures the observations, effectively flipping the standard orientation.

It is desirable that the centroids of the conditional distributions lie on the identity line. Recall that a forecast is perfectly calibrated if $E(x|f) = f$. Furthermore, a forecast is discriminatory (type 2 conditional bias) if we see predictions that have a high concentration around the specific x, $E(f|x) \approx x$. Calibration and type 2 conditional bias relate to the second terms in the CR and LB factorizations, $E_f[f - E(x|f)]^2$ and $E_x[x - E(f|x)]^2$, and can be examined visually using the conditional plots of Figure 5. The third terms of the CR and LB factorizations are best interpreted quantitatively. They relate to resolution and discrimination.

Figure 5: Conditional distributions of $p(x)$ and $p(f)$ showing the variability of x given f and the variability of f given x. Black triangles mark the values at which the centroids should lie, peaks are marked by the yellow dots.



Examining the conditional plots of Figure 5 for the variability of observed GHI given forecasted GHI, we note that generally the centroids lie close to the expected value for almost all ranges of GHI. The exception to this are the rows for forecasted GHI in ranges 100 to 200 and 200 to 300. This observation of higher error rates towards overestimation for these early GHI ranges was also part of the scatter plot analysis in Figure 3. However, beyond these specific instances, the observation that forecasts are quite well calibrated holds consistent across different temperature and wind speed buckets. This indicates a small type 1 conditional bias, suggesting that the forecasts are well calibrated throughout the GHI range and for situations of both high and low temperature and wind speeds. Looking at the specific differences between buckets, for temperature, the highest temperature bucket seems most calibrated as it doesn't feature the overestimation bias described prior. For wind speed, the centroids tend to be positioned close to the expected value for all buckets; however, it is notable that as wind increases, the density spreads out further away from the peak of the distribution, indicating increased uncertainty.

The conditional plots for variability of forecasted GHI given observed GHI reveal a noticeable deviation from the identity line for all GHI ranges except

0 to 100 that is consistent throughout different temperature and wind speed buckets. It appears that as we go towards higher and higher GHI ranges, the type 2 conditional bias worsens and the centroids of the distribution fall further and further behind the expected value. This underestimation was also visible in Figure 3, where the 2D contours showed most data points to lie below the identity line, except for very low values of GHI (<100). Comparison between the temperature buckets reveals that as we go towards lower temperatures, this shift towards underestimation for higher GHI ranges worsens. The difference between the wind speed buckets seems to stem not from the worsening of the bias as we go towards higher wind speeds but from an increase in uncertainty represented by the spread of the density away from the peak and the ticker distribution tails.

The visual analysis of the conditional distribution is summarized in that the type 1 conditional bias, $E_f[f - E(x|f)]^2$, is relatively small, meaning that the forecast is generally well calibrated. This is especially true when it comes to states of high temperature or low wind. The same holds for type 2 conditional bias, $E_x[x - E(f|x)]^2$, where the forecast was most discriminatory for states with higher temperatures and lower wind speeds.

Temp	**MSE**	$V(f)$	$V(x)$	$E_f\left[f - E(x\|f)\right]^2$	$E_f\left[E(x\|f) - E(x)\right]^2$
All Data	8732.585	46058.156	53268.157	147.812	44683.064
(-7.47 - 8.89)	4501.920	15310.078	19803.210	544.216	15863.848
(8.89 - 14.83)	7679.438	29240.461	34302.627	586.282	27236.869
(14.84 - 20.8)	11276.528	48533.337	56318.291	622.770	45636.606
(20.8 - 35.08)	11475.029	48395.830	57902.121	562.389	46923.952

Table 5: Error terms based on calibration-refinement factorization according to temperature.

Wind	**MSE**	$V(f)$	$V(x)$	$E_f\left[f - E(x\|f)\right]^2$	$E_f\left[E(x\|f) - E(x)\right]^2$
All Data	8732.585	46058.156	53268.157	147.812	44683.064
(0.42 - 2.0)	5385.520	42847.004	50147.283	491.126	45240.178
(2.01 - 3.5)	9745.503	50735.240	57037.423	829.646	48145.874
(3.5 - 5.33)	10198.629	46722.880	53813.055	620.317	44207.337
(5.33 - 11.42)	9602.726	43393.441	51687.939	544.188	42665.475

Table 6: Error terms based on calibration-refinement factorization according to wind speed.

4.7 Quantitative analysis of the conditional distribution

The quantitative summary based on the CR and LB factorizations allows us to confirm the conclusions of the previous visual analysis. Both terms, $E_f[f - E(x|f)]^2$ and $E_x[x - E(f|x)]^2$, however, feature the conditional distributions of $p(x|f)$ and $p(f|x)$ in the form of the conditional means $E(x|f)$ and $E(f|x)$. To calculate this, we employ kernel conditional density estimation (KCDE) twice since both f and x take the place of the independent and dependent variables, depending on the factorization used. For details on the KCDE used in this context, see [41].The calculation of $E(x|f)$ and $E(f|x)$ for the last terms of the factorizations is analogous.

The final aspects of forecast quality that we will examine are represented by the third terms of the CR and LB factorization, $E_f[E(x|f) - E(x)]^2$ and $E_x[E(f|x) - E(f)]^2$, which relate to resolution and discrimination, respectively. In the MSE decomposition, both terms have a negative sign, indicating that larger differences are preferred. The resolution for both temperature and wind speed is tabulated in Tables 5 and 6. A pattern emerges for the temperature bins where, as we go towards higher temperature bins, resolution significantly increases. Resolution relates to the conditional expectations and the marginal expectation, $E(x|f) - E(x)$. It is desirable for the difference between these terms to be as large as possible, as it would mean that based on the forecasts, the observations associated with them would be quite different from what was expected unconditionally (marginally). The values indicate that for states with lower temperatures, the forecast has significantly less resolution; for wind speed, generally, circumstances with low to moderate wind speed seem to offer better resolution; however, this effect is much less pronounced. Similar is the situation with the third term of the LB factorization: discrimination. It denotes how different forecasts are associated with different observation values $E_x[E(f|x) - E(f)]^2$. Larger values are best, as it would mean that based on the observations, the forecasts associated with them would be quite different from what was expected unconditionally (marginally). The discrimination for both temperature and wind speed is listed in Tables 7 and 8. Here we again note a significant increase as we go towards higher temperature buckets. States with lower to moderate wind speeds appear to be only marginally better than states with higher wind speeds.

At first glance, we see these very high resolution and discrimination values as we go towards high temperatures; however, this doesn't mean that the forecast itself exhibits this severely reduced resolution and discrimination for low temperatures. This is due to the fact that variances differ substantially between the different temperature buckets. These two qualities are much more similar for wind speed across the bins, where differences in variance are notably slighter. In fact, if we were to verify our visual analysis of the conditional distribution, recall that we concluded that the highest temperature bin had the least type 2 conditional bias. If we look at Table 7, it appears that the summary measure contradicts our observations, where the highest temperature bin has the most type 2 error. This issue would self-evidently limit the interpretation of the summary metrics relating to different forecast quality measures. To attempt to overcome that, the next section describes a post-processing step that is based on a linear correction of the variances within the temperature and wind speed buckets.

Temp	**MSE**	$V(f)$	$V(x)$	$E_x[x - E(f\|x)]^2$	$E_x[E(f\|x) - E(f)]^2$
All Data	8732.585	46058.156	53268.157	1455.785	38799.998
(-7.47 - 8.89)	4501.920	15310.078	19803.210	1363.341	12193.276
(8.89 - 14.83)	7679.438	29240.461	34302.627	1640.933	23240.020
(14.84 - 20.8)	11276.528	48533.337	56318.291	2431.183	39742.300
(20.8 - 35.08)	11475.029	48395.830	57902.121	2557.297	39524.728

Table 7: Error terms based on likelihood-base rate factorization according to temperature.

Wind	**MSE**	$V(f)$	$V(x)$	$E_x[x - E(f\|x)]^2$	$E_x[E(f\|x) - E(f)]^2$
All Data	8732.585	46058.156	53268.157	1455.785	38799.998
(0.42 - 2.0)	5385.520	42847.004	50147.283	1223.085	38691.525
(2.01 - 3.5)	9745.503	50735.240	57037.423	1780.392	42691.145
(3.5 - 5.33)	10198.629	46722.880	53813.055	2291.694	38925.561
(5.33 - 11.42)	9602.726	43393.441	51687.939	2281.822	36061.746

Table 8: Error terms based on likelihood-base rate factorization according to wind speed.

4.8 Linear variance correction

In the previous sections, the analysis was based on four equally sized bins for both wind speed and temperature; however, the variance within those buckets can be quite different, especially for the latter, where the highest temperature bucket has close to three times the variance of the lowest temperature bucket. Because of these reasons, we review a post-processing strategy from [42] where the variance of the forecasts is corrected and the MBE is eliminated.

Corrected forecasts are represented by the linear equation:

$$f^* = af + b, \tag{11}$$

where a is the slope that adjusts the scale, and b is the intercept that aligns the mean with observed values. Using the principles of statistical expectation and variance, the mean and variance of the corrected forecasts f^* are derived as

$$E(f^*) = aE(f) + b, \tag{12}$$
$$V(f^*) = a^2V(f), \tag{13}$$

respectively. Since in this correction scheme we also correct for the MBE to be zero:

$$\mathrm{MBE}(f^*, x) = E(f^*) - E(x) = aE(f) + b - E(x), \tag{14}$$

b would have to equal to $E(x) - aE(f)$. Based on the requirements of $V(f^*) = V(x)$ and $V(f^*) = a^2V(f)$ it is derived that $a = \sqrt{\frac{V(x)}{V(f)}}$. Substituting for a and b in Eq. 11 gives us the desired linear correction:

$$f^* = \sqrt{\frac{V(x)}{V(f)}}(f - E(f)) + E(x). \tag{15}$$

We have to alter the implementation of the approach above since we are not so interested in dealing with correcting the variance of the forecast in regards to observations but rather in equal variance correction of the observation and forecast variances across the stratified buckets. Below, we will only outline the forecast correction, as the observation correction is analogous.

The correction scheme follows the same linear equation from Eq. 11. The requirements here are that the variance of the bucket $V(f^*)$ would be equal to the variance across all data $V(f_{\mathrm{all}})$. Based on Eq. 13, the slope that adjusts the scale is $\sqrt{\frac{V(f_{\mathrm{all}})}{V(f)}}$. For the intercept b, we want to maintain the mean of the buckets the same, since otherwise we would lose information in regards to the interpretation of the second and third terms of the CB and LB factorizations. To fulfill this, we need $E(f^*) = E(f)$, following the mean of the corrected forecasts:

$$E(f^*) = aE(f) + b, \tag{16}$$
$$E(f) = aE(f) + b, \tag{17}$$
$$b = E(f) - aE(f). \tag{18}$$

Substituting for a and b in Eq. 11 gives us the desired linear correction:

$$f^* = \sqrt{\frac{V(f_{\text{all}})}{V(f)}}(f - E(f)) + E(f). \tag{19}$$

The results of the correction are listed in Tables 9 and 10. It is clear that after equating the variance within bins, the type 1 bias is no longer similar between different temperature states. The higher the temperature, the better calibrated the forecasts are; however, the resolution does not follow this pattern. It appears similar across the temperature range and is actually highest (by a small amount) in the lowest temperature circumstances. For wind speed, the corrected terms are close to the original for calibration; nevertheless, resolution makes the biggest difference where it is markedly higher for low wind conditions.

Comparable are also the results from the second and third terms of the likelihood-base rate factorization. Again, in contrast to the original results, as temperature increases, the type 2 conditional bias decreases. Here we further note an increase in discrimination for states with mid- to high temperatures. This increase in discrimination is more significant for the lowest wind bin, which is also the one with the least type 2 conditional bias. The recalculated general error metrics are tabulated in Tables 11 and 12. The mean bias error stays the same as our correction approach maintains the local means within the bins. The other corrected error terms, however, differ from the original and result in insight quite similar to the normalized metrics we calculated for the quantitative assessment of the joint distribution. Overall, the linear correction scheme employed aligned much more closely with the visual analysis than the originally calculated terms.

Temp	**MSE**	$E_f\,[f - E(x\vert f)]^2$	$E_f\,[E(x\vert f) - E(x)]^2$	$E_x[x - E(f\vert x)]^2$	$E_x[E(f\vert x) - E(f)]^2$
(-7.47 - 8.89)	12225.57	2291.76	43491.57	3058.98	36910.73
(8.89 - 14.83)	11926.11	1071.42	42496.17	2512.11	36690.46
(14.84 - 20.8)	10667.83	580.11	43126.86	2263.53	37676.97
(20.8 - 35.08)	10596.17	550.85	43136.14	2090.86	37578.41

Table 9: Corrected error terms based on calibration-refinement and likelihood-base rate factorization according to temperature.

Wind	**MSE**	$E_f[f - E(x\|f)]^2$	$E_f[E(x\|f) - E(x)]^2$	$E_x[x - E(f\|x)]^2$	$E_x[E(f\|x) - E(f)]^2$
(0.42 - 2.0)	5702.31	517.99	48045.97	1237.71	41576.51
(2.01 - 3.5)	9064.22	714.31	44926.47	1820.54	38726.76
(3.5 - 5.33)	10085.67	605.17	43736.20	2292.46	38351.13
(5.33 - 11.42)	9920.15	605.51	43969.23	2158.96	38267.01

Table 10: Corrected error terms based on calibration-refinement and likelihood-base rate factorization according to wind speed.

Temp	**MBE**	**cMBE**	**RMSE**	**cRMSE**	**MAE**	**cMAE**
(-7.47 - 8.89)	2.962576	2.962576	67.096348	110.569285	48.831114	78.780703
(8.89 - 14.83)	6.674911	6.674911	87.632405	109.206718	62.467249	77.811068
(14.84 - 20.8)	4.864266	4.864266	106.190996	103.285176	71.674116	69.658918
(20.8 - 35.08)	-2.328031	-2.328031	107.121563	102.937688	69.583012	66.883102

Table 11: Original vs. corrected general error metrics for different temperature bins.

Wind	**MBE**	**cMBE**	**RMSE**	**cRMSE**	**MAE**	**cMAE**
(0.42 - 2.0)	-8.782939	-8.782939	73.386101	75.513673	52.096576	53.403360
(2.01 - 3.5)	5.875879	5.875879	98.719316	95.206206	65.748681	63.933766
(3.5 - 5.33)	9.334237	9.334237	100.988260	100.427459	66.327929	66.021158
(5.33 - 11.42)	5.753688	5.753688	97.993499	99.599973	68.380319	69.322967

Table 12: Original vs. corrected general error metrics for different wind speed bins.

5 Discussion and Conclusion

A central goal of this thesis was to conduct a comprehensive analysis of forecast quality, focusing not only on general aspects but also going into specific dimensions such as calibration, type 2 conditional bias, resolution, as well as discrimination across varying weather conditions like temperature and wind speed. This was achieved by leveraging the synergy between the traditional measure-oriented approach and the more recently introduced distribution-oriented approach within the solar forecasting community. To ensure completeness, both visual and quantitative assessments were considered. In light of this, because the numerical evaluation of the distribution-oriented approach proves to be more burdensome, an alternative is to heavily lean towards visual analysis since it is not only more straightforward but also allows for a detailed evaluation of the forecast properties without the need for complex calculations or post-processing schemes.

The implementation initially focuses on the general error metrics of RMSE and MAE, given that MBE is now commonly seen more as a standard requirement than a measure of quality. This shift is due to most modern forecasting models incorporating some level of bias correction. Furthermore, compared to MBE, the visual analysis in the form of scatter plots allows for a more detailed analysis of the bias throughout the whole GHI range. When assessing these general error metrics, it is crucial to highlight the significance of normalization. Without normalization, RMSE and MAE are difficult to interpret and can obscure valuable insights, as these metrics depend on the scale of the data, which varies for different temperature and wind speed states. The variability of solar irradiance primarily depends on factors such as geographical location, seasonality, and timescale. Even when data are matched based on these factors, stratifying the data into four equal-sized bins for both temperature and wind speed results in subsets with differing variances. This hinders the interpretability of the terms derived from the calibration-refinement and likelihood-based rate factorization. To address this issue, a post-processing step involving a simple linear variance correction is implemented.

The results of the analysis revealed that the accuracy of weather forecasts indeed varies under different weather conditions. Specifically, forecasts are more accurate in higher-temperature environments compared to lower ones. For wind speed, forecasts associated with lower wind speeds demonstrated better performance. Moreover, the quality of the forecasts displayed distinct nuances; higher temperatures were associated with better calibration and notably lower type 2 conditional bias. In contrast, lower wind speeds were found to have enhanced discrimination and higher resolution in the forecasts.

References

[1] International Energy Agency. *Renewables 2023: Analysis and Forecasts to 2028.* Organisation for Economic Co-operation and Development Publishing, 2023.

[2] Carlos Algora and Ignacio Rey-Stolle. *Handbook of concentrator photovoltaic technology.* John Wiley & Sons, 2016.

[3] Amine Allouhi, Shafiqur Rehman, Mahmut Sami Buker, and Zafar Said. Up-to-date literature review on solar pv systems: Technology progress, market status and r&d. *Journal of Cleaner Production*, 362:132339, 2022.

[4] J Antonanzas, D Pozo-Vázquez, LA Fernandez-Jimenez, and FJ Martinez-de Pison. The value of day-ahead forecasting for photovoltaics in the spanish electricity market. *Solar Energy*, 158:140–146, 2017.

[5] G. Austria. Messstationen zehnminutendaten v2, 2024. [Data set].

[6] B Brown, F Atger, H Brooks, B Casati, U Damrath, B Ebert, A Ghelli, P Nurmi, D Stephenson, C Wilson, et al. Recommendations for the verification and intercomparison of qpfs from operational nwp models. *World Meteorological Organization: Geneva, Switzerland*, 2004.

[7] B Casati, LJ Wilson, DB Stephenson, P Nurmi, A Ghelli, M Pocernich, U Damrath, EE Ebert, BG Brown, and S Mason. Forecast verification: current status and future directions. *Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling*, 15(1):3–18, 2008.

[8] Romain Dambreville, Philippe Blanc, Jocelyn Chanussot, and Didier Boldo. Very short term forecasting of the global horizontal irradiance using a spatio-temporal autoregressive model. *Renewable Energy*, 72:291–300, 2014.

[9] Burcin Cakir Erdener, Cong Feng, Kate Doubleday, Anthony Florita, and Bri-Mathias Hodge. A review of behind-the-meter solar forecasting. *Renewable and Sustainable Energy Reviews*, 160:112224, 2022.

[10] Thomas E Hoff, Richard Perez, Jan Kleissl, David Renne, and Joshua Stein. Reporting of irradiance modeling relative prediction errors. *Progress in Photovoltaics: Research and Applications*, 21(7):1514–1519, 2013.

[11] Tao Hong and Shu Fan. Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting*, 32(3):914–938, 2016.

[12] Tao Hong, Pierre Pinson, Yi Wang, Rafał Weron, Dazhi Yang, and Hamidreza Zareipour. Energy forecasting: A review and outlook. *IEEE Open Access Journal of Power and Energy*, 7:376–388, 2020.

[13] Jing Huang, Małgorzata Korolkiewicz, Manju Agrawal, and John Boland. Forecasting solar radiation on an hourly time scale using a coupled autoregressive and dynamical system (cards) model. *Solar Energy*, 87:136–149, 2013.

[14] Xiaoqiao Huang, Qiong Li, Yonghang Tai, Zaiqing Chen, Jun Zhang, Junsheng Shi, Bixuan Gao, and Wuming Liu. Hybrid deep neural model for hourly solar irradiance forecasting. *Renewable Energy*, 171:1041–1060, 2021.

[15] Dongyu Jia, Jiajia Hua, Liping Wang, Yitao Guo, Hong Guo, Pingping Wu, Min Liu, and Liwei Yang. Estimations of global horizontal irradiance and direct normal irradiance by using fengyun-4a satellite data in northern china. *Remote Sensing*, 13(4):790, 2021.

[16] Ian T Jolliffe and David B Stephenson. *Forecast verification: a practitioner's guide in atmospheric science.* John Wiley & Sons, 2012.

[17] Dhivya Sampath Kumar, Gokhan Mert Yagli, Monika Kashyap, and Dipti Srinivasan. Solar irradiance resource and forecasting: a comprehensive review. *IET Renewable Power Generation*, 14(10):1641–1656, 2020.

[18] Pratima Kumari and Durga Toshniwal. Deep learning models for solar irradiance forecasting: A comprehensive review. *Journal of Cleaner Production*, 318:128566, 2021.

[19] V Lara-Fanego, JA Ruiz-Arias, D Pozo-Vázquez, FJ Santos-Alamillos, and J Tovar-Pescador. Evaluation of the wrf model solar irradiance forecasts in andalusia (southern spain). *Solar Energy*, 86(8):2200–2217, 2012.

[20] Philippe Lauret, Mathieu David, and Pierre Pinson. Verification of solar irradiance probabilistic forecasts. *Solar Energy*, 194:254–271, 2019.

[21] Philippe Lauret, Cyril Voyant, Ted Soubdhan, Mathieu David, and Philippe Poggi. A benchmarking of machine learning techniques for solar radiation forecasting in an insular context. *Solar Energy*, 112:446–457, 2015.

[22] Edward W Law, Abhnil A Prasad, Merlinde Kay, and Robert A Taylor. Direct normal irradiance forecasting and its application to concentrated solar thermal output forecasting–a review. *Solar Energy*, 108:287–307, 2014.

[23] Ricardo Marquez and Carlos FM Coimbra. Forecasting of global and direct solar irradiance using stochastic learning methods, ground experiments and the nws database. *Solar Energy*, 85(5):746–756, 2011.

[24] Jonathan R Moskaitis. A case study of deterministic forecast verification: Tropical cyclone intensity. *Weather and forecasting*, 23(6):1195–1220, 2008.

[25] Allan H Murphy. What is a good forecast? an essay on the nature of goodness in weather forecasting. *Weather and forecasting*, 8(2):281–293, 1993.

[26] Allan H Murphy, Barbara G Brown, and Yin-Sheng Chen. Diagnostic verification of temperature forecasts. *Weather and Forecasting*, 4(4):485–501, 1989.

[27] Allan H Murphy and Robert L Winkler. A general framework for forecast verification. *Monthly weather review*, 115(7):1330–1338, 1987.

[28] Hugo TC Pedro and Carlos FM Coimbra. Nearest-neighbor methodology for prediction of intra-hour global horizontal and direct normal irradiances. *Renewable Energy*, 80:770–782, 2015.

[29] Sanjay Kumar Prajapati and Kishan Bhushan Sahay. A survey paper on solar irradiance forecasting methods. *Int. J. Eng. Res*, 5(03):536–541, 2016.

[30] REN21. *Renewables 2023 Global Status Report Collection*. REN21, 2023.

[31] Md Shafiullah, Shakir D Ahmed, and Fahad A Al-Sulaiman. Grid integration challenges and solution strategies for solar pv systems: a review. *IEEE Access*, 10:52233–52257, 2022.

[32] Henry R Stanski, Laurence J Wilson, and William R Burrows. Survey of common verification methods in meteorology. 1989.

[33] Yudong Tian, Grey S Nearing, Christa D Peters-Lidard, Kenneth W Harrison, and Ling Tang. Performance metrics, error modeling, and uncertainty quantification. *Monthly Weather Review*, 144(2):607–613, 2016.

[34] Loïc Vallance, Bruno Charbonnier, Nicolas Paul, Stéphanie Dubost, and Philippe Blanc. Towards a standardized procedure to assess solar forecast accuracy: A new ramp and time alignment metric. *Solar Energy*, 150:408–422, 2017.

[35] Marta Victoria, Nancy Haegel, Ian Marius Peters, Ron Sinton, Arnulf Jäger-Waldau, Carlos del Canizo, Christian Breyer, Matthew Stocks,

Andrew Blakers, Izumi Kaizuka, et al. Solar photovoltaics is ready to power a sustainable future. *Joule*, 5(5):1041–1056, 2021.

[36] Daniel S Wilks. *Statistical methods in the atmospheric sciences*, volume 100. Academic press, 2011.

[37] Dazhi Yang. A correct validation of the national solar radiation data base (nsrdb). *Renewable and Sustainable Energy Reviews*, 97:152–155, 2018.

[38] Dazhi Yang, Stefano Alessandrini, Javier Antonanzas, Fernando Antonanzas-Torres, Viorel Badescu, Hans Georg Beyer, Robert Blaga, John Boland, Jamie M Bright, Carlos FM Coimbra, et al. Verification of deterministic solar forecasts. *Solar Energy*, 210:20–37, 2020.

[39] Dazhi Yang and Jamie M Bright. Worldwide validation of 8 satellite-derived and reanalysis solar radiation products: A preliminary evaluation and overall metrics for hourly data over 27 years. *Solar Energy*, 210:3–19, 2020.

[40] Dazhi Yang, Jan Kleissl, Christian A Gueymard, Hugo TC Pedro, and Carlos FM Coimbra. History and trends in solar irradiance and pv power forecasting: A preliminary assessment and review using text mining. *Solar Energy*, 168:60–101, 2018.

[41] Dazhi Yang and Richard Perez. Can we gauge forecasts using satellite-derived solar irradiance? *Journal of Renewable and Sustainable Energy*, 11(2), 2019.

[42] Dazhi Yang, Wenting Wang, Jamie M Bright, Cyril Voyant, Gilles Notton, Gang Zhang, and Chao Lyu. Verifying operational intra-day solar forecasts from ecmwf and noaa. *Solar Energy*, 236:743–755, 2022.

[43] Dazhi Yang, Wenting Wang, Christian A Gueymard, Tao Hong, Jan Kleissl, Jing Huang, Marc J Perez, Richard Perez, Jamie M Bright, Xiang'ao Xia, et al. A review of solar forecasting, its dependence on atmospheric sciences and implications for grid integration: Towards carbon neutrality. *Renewable and Sustainable Energy Reviews*, 161:112348, 2022.

[44] Jie Zhang, Anthony Florita, Bri-Mathias Hodge, Siyuan Lu, Hendrik F Hamann, Venkat Banunarayanan, and Anna M Brockway. A suite of metrics for assessing the performance of solar power forecasting. *Solar Energy*, 111:157–175, 2015.