
Contents lists available at SciVerse ScienceDirect
Progress in Energy and Combustion Science
journal homepage: www.elsevier.com/locate/pecs

Review
Solar forecasting methods for renewable energy integration
Rich H. Inman, Hugo T.C. Pedro, Carlos F.M. Coimbra*
Department of Mechanical and Aerospace Engineering, Jacobs School of Engineering, Center of Excellence in Renewable Energy Integration and Center for Energy Research, University of California, 9500 Gilman Drive, La Jolla, CA 92037, USA

ARTICLE INFO
Article history:
Received 14 January 2013
Accepted 16 June 2013
Available online 26 July 2013

Keywords:
Weather-dependent renewable energy
Solar forecasting
Solar meteorology
Solar variability
Solar energy integration
Evolutionary forecasting methods

ABSTRACT
The higher penetration of renewable resources in the energy portfolios of several communities accentuates the need for accurate forecasting of variable resources (solar, wind, tidal) at several different temporal scales in order to achieve power grid balance. Solar generation technologies have experienced strong energy market growth in the past few years, with corresponding increase in local grid penetration rates. As is the case with wind, the solar resource at the ground level is highly variable mostly due to cloud cover variability, atmospheric aerosol levels, and indirectly and to a lesser extent, participating gases in the atmosphere. The inherent variability of solar generation at higher grid penetration levels poses problems associated with the cost of reserves, dispatchable and ancillary generation, and grid reliability in general. As a result, high accuracy forecast systems are required for multiple time horizons that are associated with regulation, dispatching, scheduling and unit commitment. Here we review the theory behind these forecasting methodologies, and a number of successful applications of solar forecasting methods for both the solar resource and the power output of solar plants at the utility scale level.

© 2013 Elsevier Ltd. All rights reserved.

Contents
1. Introduction .......................................................... 536
2. Fundamental considerations ...................................... 537
  2.1. Solar irradiance components ................................... 537
  2.2. Airmass and Linke Turbidity ................................... 537
  2.3. Clear sky models ................................................. 537
    2.3.1. Solis model .................................................. 537
    2.3.2. European Solar Radiation Atlas (ESRA) model .......... 537
    2.3.3. Bird and Hulstrom model ................................. 538
    2.3.4. Molineaux model ........................................... 538
    2.3.5. Ineichen model ............................................. 538
    2.3.6. CPCR2 model .............................................. 538
    2.3.7. REST2 model ............................................... 538
    2.3.8. Kasten model .............................................. 538
    2.3.9. Polynomial fit ............................................... 538
    2.3.10. ASCE evapotranspiration model .......................... 538
  2.4. Clear sky & clearness indices ................................ 538
  2.5. Persistence forecasts ........................................... 538
  2.6. Evaluation of solar forecasting skill ........................ 539
    2.6.1. Solar variability .......................................... 540
    2.6.2. Forecast uncertainty ....................................... 540
    2.6.3. Performance metric ........................................ 541
3. Regressive methods ................................................ 541
  3.1. Linear stationary models ...................................... 543
    3.1.1. Auto-Regressive (AR) models ............................. 543

* Corresponding author. Tel.: +1 858 534 4285.
E-mail address: ccoimbra@ucsd.edu (C.F.M. Coimbra).

0360-1285/© 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.pecs.2013.06.002
3.12. Moving Average (MA) models ......................................................... 543
3.13. Mixed Auto-Regressive Moving Average (ARMA) models .......... 544
3.14. Mixed Auto-Regressive Moving Average models with exogenous variables (ARMAX) ........................................ 544
3.2. Non-linear stationary models ......................................................... 545
3.3. Linear non-stationary models ......................................................... 545
3.3.1. Auto-Regressive Integrated Moving Average (ARIMA) models .............................................................. 545
3.3.2. Auto-Regressive Integrated Moving Average models with exogenous variables (ARIMAX) .......................................... 546
4. Artificial Intelligence (AI) techniques ............................................... 546
4.1. Artificial Neural Networks (ANNs) ................................................. 547
4.1.1. Early networks ........................................................................... 548
4.1.2. Simple perceptron ..................................................................... 548
4.1.3. Adaptive Linear Neuron (ADALINE) ........................................ 550
4.4. Applications of ANNs ................................................................. 553
4.5. k-nearest neighbors (k-NN) ......................................................... 553
5. Remote sensing models ............................................................... 554
5.1. Orbits ....................................................................................... 555
5.2. Satellites .................................................................................. 555
5.3. Radiation budget ........................................................................ 556
5.4. Physical satellite models ............................................................. 557
5.4.1. Gautier-Dialk-Masse model .................................................... 557
5.4.2. Marullo-Dalu-Viola model ...................................................... 557
5.4.3. Möser-Raschke model ............................................................ 557
5.4.4. Dedieu-Deschamps-Kerr model ............................................ 558
5.5. Statistical satellite models ............................................................. 558
5.5.1. Hay-Hanson model ................................................................. 558
5.5.2. Tarpley & Justus-Paris-Tarpley models .................................... 558
5.5.3. Cano-HELIOSAT model ........................................................ 559
5.5.4. Perez Operational model ......................................................... 559
6. Numerical Weather Prediction (NWP) ............................................. 561
6.1. Global Forecast System (GFS) ....................................................... 561
6.2. Regional NWP models ............................................................... 561
6.2.1. Rapid Update Cycle (RUC)/RAPid refresh (RAP) models ...... 561
6.2.2. North American Mesoscale (NAM) model ................................ 561
6.2.3. High Resolution Rapid Refresh (HRRR) model ...................... 561
6.2.4. Weather Research and Forecasting (WRF) model ................. 561
7. Local sensing .............................................................................. 562
7.1. Sky imagers .............................................................................. 562
7.2. Wireless sensor networks ............................................................ 563
7.3. Pyranometer arrays .................................................................... 563
8. Hybrid systems ........................................................................... 564
9. Performance of forecasting techniques .......................................... 565
9.1. NWP-based forecast ................................................................. 565
9.2. Stochastic forecasts ................................................................... 568
9.3. AI forecasting models .............................................................. 568
9.4. Hybrid forecasting models ........................................................ 569
10. Other challenges of solar integration ............................................ 569
11. Conclusion ................................................................................ 570

References ..................................................................................... 571

1. Introduction

The global demand for renewable energy integration to the power grids highlights the importance of economic and technological issues associated with growing levels of flat-panel Photovoltaic (PV), Concentrated Solar Power (CSP), and Concentrated PV (CPV) penetrations into the power grid. These concerns arise from the variable nature of the solar resource, seasonal deviations in production and load profiles, the high cost of energy storage, and the balance between grid flexibility and reliability [1,2]. As a result, solar plants are often backed by ancillary generators for periods of high variability, which increases the capital and operational costs of solar generation. Accurate solar forecasts over several time horizons are required so that Independent System Operators (ISOs) or equivalent grid balancing authorities are able to successfully integrate increased levels of solar power production while maintaining reliability [3–5]. Solar forecasts on multiple time horizons become increasingly important as solar penetration grows for the purposes of grid regulation, load-following production, power scheduling and unit commitment. Short-term, intra-hour solar forecasts are particularly useful for power plant operations, grid balancing, real-time unit dispatching, automatic generation control (AGC) and trading. Forecasts for longer time horizons are of interest to utilities and ISOs for unit commitment, scheduling and for improving balance area control performance. Ultimately, a spectrum of solar forecasts is required to address the planning,
operational and balancing needs of both the distribution and the transmission grids. Solar forecasting is therefore an enabling technology for the integration of ever increasing level of solar penetration into the grid because it improves the quality of the energy delivered to the grid and reduces the ancillary costs associated with weather dependency. The combination of these two factors (better energy quality through information that is capable of lowering integration and operational costs) has been the driving motivation for the development of a complex field of research that aims at producing better solar forecasting capabilities for the solar resource at the ground level and for the power output from different solar technologies that depend on the variable irradiance at the ground level. Solar, wind and load forecasting have become integral parts of the so-called ‘smart grid concept’.

To date, high-fidelity, robust solar forecast systems that work for widely different microclimates remain evasive. The problem is of great complexity due to the non-linear and chaotic effect of cloud motion on solar irradiance at the ground level. However, a number of promising approaches have been developed in the past few years, and the incipient research field of solar meteorology for renewable generation has grown considerably by aggregating diverse areas of knowledge such as atmospheric physics, solar instrumentation, machine learning, forecasting theory and remote sensing in its quest to better predictive skills. This work presents an overview of the forecasting methods for solar resourcing and solar power generation, as well as the theoretical basis for the most promising methods, and a discussion on their effectiveness for operational use.

2. Fundamental considerations

Here we review some fundamental concepts and methods including solar irradiance components, airmass, Linke turbidity, clear sky models, clear sky and clearness indices, persistence forecasts, and a metric for the evaluation of solar forecasting skill. These topics lay the foundation for the development of solar forecasting methodologies that are covered in the following sections.

2.1. Solar irradiance components

The total extraterrestrial beam irradiance incident on the earth’s atmosphere $I_0$ fluctuates about an average value of approximately 1360 Wm$^{-2}$ [6]. This incident radiation is attenuated as it negotiates its way to the ground level through a complex series of multiple reflections, absorptions and re-emissions due to interactions with atmospheric constituents [7]. This results in the division of the incident extraterrestrial beam radiation into two distinct components: Direct Normal Irradiance (DNI) and Diffuse Horizontal Irradiance (DHI). The geometric sum of which is the Global Horizontal Irradiance (GHI) which can be written as

$$\text{GHI} = \text{DHI} + \text{DNI} \cdot \cos \theta$$

(1)

where $\theta$ is the solar zenith angle.

2.2. Airmass and Linke Turbidity

Two coefficients which are employed quite often in solar engineering are the atmospheric airmass (AM) and the Linke Turbidity ($T_L$). In the context of this review, airmass refers to the optical path length of the atmospheric boundary layer through which radiation must propagate in order to reach ground level. When the sun is directly overhead, $\theta = 0$, the airmass has a value of unity, AM = 1. As the solar zenith angle $\theta$ increases, so does the airmass. Many models of the airmass as a function of solar zenith angle, AM($\theta$), exist (see Ref. [8] for a more detailed review). It is convenient to note that the irradiance at AM = 0 is the extraterrestrial irradiance $I_0 = 1360$ Wm$^{-2}$ [6].

The Linke turbidity coefficient represents the number of clean and dry atmospheres that would be required to produce the observed irradiance at ground level. In 1922 Linke proposed that the total optical thickness of a cloudless atmosphere could be expressed as a function of the optical thickness of a clear and dry atmosphere, $\delta_{dca}$, and the Linke turbidity coefficient, $T_L$, as

$$\text{DNI} = I_0 \exp(-\delta_{dca} - T_L \cdot \text{AM}).$$

(2)

It should be noted that several improvements have been made to this model including Kasten and Young’s 1989 correction [9] and Ineichen’s airmass independent formulation [10]. Again, the reader is referred to [8] for more information.

2.3. Clear sky models

Many solar irradiance forecasting models require the knowledge of clear sky conditions. Clear sky conditions are characterized by the absence of clouds. Clear sky models are used for persistence forecasts, discussed in Section 2.5, and for normalization of the forecasting skill metric described in Section 2.6. Satellite based forecasts, covered in Section 5, also use clear sky models to derive ground solar irradiance components through normalization of satellite data and are frequently used for locations where no other measurements exist.

Clear sky models are typically developed using one of several Radiative Transfer Models (RTMs) and require local meteorological inputs such as ozone content, water vapor content and/or Linke turbidity in combination with solar geometry. A comprehensive comparison of eight clear sky models against 16 independent data banks was published by Ineichen in 2006 [11]. Ineichen concluded that the input parameters, most notably turbidity, had the highest influence on model accuracy. In addition, Ineichen found that accuracy is not highly dependent on the model itself. Therefore, the model selection criteria can be tailored to the application and availability of data. If access to comprehensive climatic input parameters are not available, the ESRA or Ineichen model would be appropriate due to the requirement of only a single input parameter. If access to input parameters is not an issue or if spectral resolution is required, the Solis model would be an appropriate choice as it gives the overall best results according to [11]. The next section briefly describes ten different clear sky models, eight of which were included in Ineichen’s comprehensive comparison.

2.3.1. Solis model

The Solis model is a spectrally resolved clear sky transmittance model that was developed by Mueller et al. [12] in 2004 for the European Heliosat-3 project. The Solis model makes use of an RTM, Beer–Lambert functions and integration over the solar spectrum in order to obtain irradiance components. Necessary inputs include ozone content, water vapor content and aerosol optical depth at 550 nm.

2.3.2. European Solar Radiation Atlas (ESRA) model

The European Solar Radiation Atlas (ESRA) model was developed for the Heliosat-2 project through the work of Rigollier et al. [13] and refined by Geiger et al. [14]. ESRA was later adapted by Zarzalejo et al. [15] and Badescu [16]. The model is derived from the Linke turbidity factor and Kasten’s Rayleigh optical depth parameterization [17]. The only required input parameter is the Linke turbidity factor at an airmass of 2.
2.3.3. Bird and Hulstrom model

As the name implies, Bird and Hulstrom [18] developed this model using RTM schemes based on McClatchey and Selby’s LOWTRAN [19]. The authors referred to this new scheme as SOLTRAN. Three parameters are required as inputs: water vapor column, aerosol optical depth (at 380 and 500 nm) and the ozone column. For more information see Refs. [11,18].

2.3.4. Molineaux model

The Molineaux model was developed in 1998 by Molineaux [20] and is based on a Moderate Resolution model using RTM (MODTRAN) developed by Berk et al. [21] and Smarts2 which is a product of Gueymard’s work [22]. Because Molineaux’s model is based on the equivalence of pyrheliometric and mono-chromatic aerosol optical depths, it cannot be used to calculate GHI. Two inputs are required: the aerosol optical depth at 700 nm and the water vapor column.

2.3.5. Ineichen model

Previous to the development of this model, the Linke turbidity coefficient had the disadvantage of being dependent on airmass. This widely used clear sky model was developed by Ineichen and Perez in 2002 [10] to establish the Linke turbidity coefficient independently of the airmass. The model possesses the attractive property that only the Linke turbidity is required as an input.

2.3.6. CPC2 model

The CPC2 model was developed by Gueymard in 1989 [23] and is a dual band technique for modeling solar radiation which parameterizes the transmittance of each extinction layer. The components of solar radiation (DNI and GHI) are then calculated as functions of the extinction layer transmittances. Inputs to the model include the Ångstrom size coefficient, the Ångstrom turbidity coefficient and the aerosol optical depth.

2.3.7. REST2 model

REST2 is a dual-band model developed by Gueymard in 2004 [24] which is essentially a modified version of the CPC2 model. The REST2 incorporates the latest extraterrestrial spectral distribution, solar constant value and updated versions of the transmittances functions in the CPC2 model. This model requires water vapor content, Ångstrom turbidity coefficient and reduced NO2 and O3 vertical path lengths as inputs.

2.3.8. Kasten model

This model is based on Kasten’s pyrheliometric formula developed in Ref. [25]. The model calculates ground irradiance components by considering the absorption and scattering that takes place at altitudes of 8000 and 2500 m. Like the Ineichen model, it conveniently only requires the Linke turbidity at airmass 2 as an input. A full description can be found in Kasten’s 1984 work [26].

2.3.9. Polynomial fit

One of the simplest clear sky models is a location specific polynomial of the cosine of the solar zenith angle \( \cos(\theta) \) given by

\[
I_{F}^{clr} = \sum_{n=0}^{N} c_n (\cos(\theta))^n. \tag{3}
\]

Coimbra and Marquez [27] showed that typically a third order polynomial fit \((N = 3)\) is sufficient. The third order polynomial can be written

\[
I_{F}^{clr} = c_3 (\cos \theta_l)^3 + c_2 (\cos \theta_l)^2 + c_1 (\cos \theta_l) + c_0, \tag{4}
\]

where the coefficients are determined by fitting the polynomial in (4) to the clear sky days in the data set. The limitation of this model is that a historical database of clear days is required for the specific location. Typically such a historical database is not available. If this is the case, one of the other clear sky models should be employed.

2.3.10. ASCE evapotranspiration model

R. L. Snyder [28] described the steps required to estimate reference evapotranspiration using hourly weather data in a method that was adopted by the American Society of Civil Engineers (ASCE-EWRI, 2004). The method, which employs the Penman-Monteith (hourly) reference evapotranspiration equations, calculates, as an intermediate step, clear sky total global solar radiation at the Earth’s surface. This model combines Duffie and Beckman’s equations [29] for extraterrestrial radiation and a clear sky GHI model from Allen et al. [30]. Only location (latitude, longitude and altitude) is required as an input.

2.4. Clear sky & clearness indices

Two parameters which are frequently used in solar forecasting are the clear sky index and the clearness index. They are similar and differ only in the normalization factor used in their definition. The clear sky index \( k_t \) is defined as the ratio of the measured irradiance to the modeled clear sky irradiance at ground level

\[
k_t = \frac{I_t}{I_{F}^{clr}}. \tag{5}
\]

The clearness index \( K_t \) is similarly defined with normalization of the ground level irradiance with respect to the extraterrestrial irradiance

\[
K_t = \frac{I_t}{I_{F}^{EX}}. \tag{6}
\]

Extraterrestrial irradiance is much easier to model than the clear sky irradiance due to the absence of physical atmospheric fluctuations. A simple extraterrestrial irradiance model can be written as

\[
I_{F}^{EX} = I_0 \cos(\theta_t), \tag{7}
\]

where \( I_0 = 1360 \text{ W/m}^2 \) [6] is the solar constant and \( \theta_t \) is the solar zenith angle at time \( t \). In contrast to the clear sky models, this extraterrestrial irradiance model does not involve any location specific fitting parameters except for \( \theta_t \) which depends on latitude. Because \( k_t \) and \( K_t \) are dependent on \( I_{F}^{clr} \) and \( I_{F}^{EX} \) respectively, there can be several clear sky or clearness indices depending on which clear sky model or extraterrestrial model is used. Fig. 1 illustrates the relationship between the clear sky and clearness indices. It is important to note that the clear sky index \( k_t \) is always greater than the clearness index \( K_t \) due to the difference in normalization factors used.

2.5. Persistence forecasts

Persistence forecasts are some of the most simple forecasts to implement and are often used as a baseline for the performance evaluation of other forecast engines. Persistence forecasts, as the name implies, are defined as having the clear sky or clearness index persist for the next time step. For the clear sky persistence forecast we define the following persistence condition

\[
k_t + \Delta t = k_t = \frac{I_t}{I_{F}^{clr}}. \tag{8}
\]
where $\Delta t$ is the forecast horizon. Similarly, for the clearness persistence forecast we define a corresponding persistence condition

\[ K_{t+\Delta t} = K_t = \frac{I_t}{I_{clr}} \]  

(9)

From these indices, the forecast of the solar irradiance at the next time step is computed as

\[ I_{t+\Delta t} = I_t^c_{t+\Delta t} = \frac{K_{t+\Delta t}}{K_{t+\Delta t}} \]  

(10)

and

\[ I_{t+\Delta t} = I_{clr_{t+\Delta t}} \]  

(11)

Again, since these forecasts rely on the definition of a clear sky or clearness index, there can be several persistence forecasts depending on how $I_t^c$ and $I_{clr}$ are estimated. Two noteworthy properties of the persistence forecasts are shown in Fig. 2. First, during cloudless periods the persistence forecast performs quite well, however, large forecasting errors occur during rapid changes in irradiance $I$ which result from the passing of opaque clouds. Secondly, the persistence forecast displays an obvious time delay due to the implied persistence of the clear sky indices.

### 2.6. Evaluation of solar forecasting skill

Historically, traditional power generation technologies such as fossil and nuclear power which were designed to run in stable output modes, have resulted in the majority of power grid variability originating from demand fluctuations [3,27]. On the other hand, energy sources that rely on the solar resource exhibit a high degree of variability (see Fig. 3) which introduces fluctuations also on the energy supply side. Recent studies suggest that in order to facilitate higher market and grid penetration of solar power, ISOs need accurate forecasts of solar irradiance on multiple time horizons [3–5]. Despite the large number of forecasting methods described in this work, the comparison of results and evaluation of relative advantages between models has been elusive. These difficulties arise from the fact that solar irradiance is inherently dependent on geographic location, time of year and climate as well as the different evaluation techniques used by various authors for the quality assessment of their models and forecasts, denoted by $I$. Traditional statistical metrics to characterize model quality include the coefficient of determination which compares the variance of the errors to the variance of the modeled data

\[
R^2 = 1 - \frac{\text{Var}(I - \hat{I})}{\text{Var}(I)}.
\]

(12)

the Mean Absolute Error (MAE) which considers only the absolute value of the errors

\[
\text{MAE} = \frac{1}{N} \sum_{t=1}^{N} |I_t - \hat{I}_t|.
\]

(13)

the Mean Absolute Percentage Error (MAPE)

\[
\text{MAPE} = \frac{100 \%}{N} \sum_{t=1}^{N} \left| \frac{I_t - \hat{I}_t}{I_t} \right|.
\]

(14)

![Fig. 1. Time-series of measured global horizontal irradiance values $I_t$, estimated clear sky irradiance $I_t^c$, estimated extraterrestrial irradiance $I_t^{ex}$ with corresponding clear sky index $K_t$ and clearness index $K_{clr}$. Note that the clearness index is always less than the clear sky index by definition. The Ineichen model was used for $I_t^c$ and Equation (7) was used for $I_t^{ex}$ Measured hourly data is for December 13–15, 2010 at UCSD EBUII in La Jolla, CA (courtesy of Jan Kleissl). Night values have been removed for clarity.](image)

![Fig. 2. Example of persistence forecast performance for a clear and cloudy day (Oct. 11, 2010 and Oct. 13, 2010) at UCSD EBUIII in La Jolla, CA (data courtesy of Jan Kleissl). The persistence forecast performs relatively well on the clear day. However, large errors which occur during abrupt changes in measured irradiance and a ‘time delay’ are visible in the persistence forecast during the cloudy day. Clear sky index $K_t$ and absolute error $e_{abs} = |I_t - \hat{I}_t|/I_t$ of the persistence forecast are shown on the lower part of the graph for the same time steps.](image)

![Fig. 3. Measured GHI values at 5 min resolution, Ineichen model clear sky irradiance values and the calculated values of step-changes of the clear sky index $\Delta K_t$. (Data for Sep. 28–30, 2010 at EBUII in La Jolla, CA courtesy of Jan Kleissl). During clear periods the solar variability is near zero. Ramps in the solar irradiance coincide with ramps in the $\Delta K_t$ signal.](image)
and the correlation coefficient

$$\rho = \frac{\text{Cov}(\hat{I}, I)}{\text{Var}(\hat{I}) \text{Var}(I)}.$$  \hspace{1cm} (15)

In recent years, the European and International Energy Agency (IEA) developed recommendations for reporting irradiance model accuracy based on three validation metrics: Root Mean Square Error (RMSE), Mean Bias Error (MBE), and Kolmogorov–Smirnov Integral (KSI) [31–34]. The RMSE is a measure of the average spread of the errors

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{t=1}^{N} (\hat{I}_t - I_t)^2},$$  \hspace{1cm} (16)

the MBE is a measure of the average bias of the model

$$\text{MBE} = \frac{1}{N} \sum_{t=1}^{N} (\hat{I}_t - I_t),$$  \hspace{1cm} (17)

and the KSI aims to quantify the ability of a model to reproduce observed statistical distributions

$$\text{KSI} = \frac{x_{\text{min}}}{D_\text{c}} = \int_{x_{\text{min}}}^{x_{\text{max}}} D_\text{c} \, dx,$$  \hspace{1cm} (18)

where $D_\text{c}$ is the difference between two cumulative distribution functions. Trapezoidal integration is typically employed as a result of $D_\text{c}$ being a discrete variable. Respectively, these metrics provide a measure of the model’s dispersion (RMSE), average bias (MBE), and ability to reproduce frequency distributions (KSI) [35].

However, in many contexts relative error is more commonly desired than absolute error, especially for users in the utility industry. Simplified reporting approaches for relative (percent) versions RMSE and MBE were proposed in a recent paper by Hoff et al. [36]. Additionally a method for reporting %KSI was proposed by Perez et al. [36].

According to Hoff et al. [35], the MAE normalized by the average value of the irradiance provides the best practical measure of relative dispersion error based on the evaluation criteria used in their study. This measure can be written as

$$\text{MAE/Avg.} = \frac{\frac{1}{N} \sum_{t=1}^{N} |\hat{I}_t - I_t|}{\frac{1}{N} \sum_{t=1}^{N} I_t} = \left( \frac{1}{\sum_{t=1}^{N} I_t} \right) \sum_{t=1}^{N} |\hat{I}_t - I_t|$$  \hspace{1cm} (19)

and is independent of the number of observations as well as being simple to understand. Hoff et al. also found that normalization of the RMSE by the maximum nominal irradiance, or capacity $C = 1000 \text{ Wm}^{-2}$, which can be written as

$$\text{RMSE/Capacity} = \left( \frac{1}{C \sqrt{N}} \right) \sum_{t=1}^{N} (\hat{I}_t - I_t)^2$$  \hspace{1cm} (20)

is also desirable and currently widely accepted in the wind power industry.

Normalization of the KSI is typically defined as

$$\text{KSI} = \frac{x_{\text{min}}}{a_c} \int_{x_{\text{min}}}^{x_{\text{max}}} \frac{D_\text{c}}{a_c} \, dx,$$  \hspace{1cm} (21)

where the critical area $a_c$ is calculated as

$$a_c = V_c (\Delta x_{\text{max}} - x_{\text{min}})$$  \hspace{1cm} (22)

and the critical value $V_c$ is calculated for a 90% level of confidence as

$$V_c = \frac{1.63}{\sqrt{N}}, \quad N \geq 35$$  \hspace{1cm} (23)

according to [37]. In order to circumvent the strong dependence of $V_c$ on the number of samples, Perez et al. [36] suggest using the %KSI or absolute integrated difference between the modeled and measured normalized cumulative distribution functions.

More recently, Marquez and Coimbra presented a novel approach for the evaluation of the quality of forecast models based on the comparison of solar resource variability and forecast uncertainty [27]. This relationship between variability and uncertainty provides a consistent metric that is independent of the time horizon. However, the metric requires a clear sky model for normalization purposes. As discussed in Section 2.3, the accuracy of clear sky models is not dependent upon the model itself, but rather the selection of input parameters the model is dependent upon [10]. Assuming that access to accurate input parameters for clear sky modeling is available, we now address the two quantities used to determine the solar forecasting metric.

### 2.6.1. Solar variability

Variability of solar irradiance at ground level is affected by a number of factors, but most importantly solar position and cloud cover [27]. Fluctuations due to solar position are completely deterministic and are typically calculated through the use of clear sky or clearness models, see Section 2.3. Cloud induced fluctuations, on the other hand, are stochastic processes difficult to model and predict. As a result, the accuracy of the solar irradiance forecasting models depends almost exclusively on the ability to forecast the stochastic component induced by cloud cover [5,38,39]. For this reason, in the definition of solar irradiance variability, the authors in Ref. [27] addressed only the stochastic component and removed fluctuations due to annual and diurnal changes of solar position. This was accomplished by defining solar irradiance variability $V$ as the standard deviation of the step-changes of the clear sky or clearness index as

$$V = \sqrt{\frac{1}{N} \sum_{t=1}^{N} \left( \frac{I_t}{I_{t-1}} - 1 \right)^2} = \sqrt{\frac{1}{N} \sum_{t=1}^{N} (\Delta I_t)^2}.$$  \hspace{1cm} (24)

The definition of solar variability above is essentially synonymous with those used by Kleissl and Lave [40] and Hoff [41] except for the refinement to include the deterministic changes as done in Refs. [38,39]. Fig. 3 shows a sequence of clear and cloudy days with the associated values of $\Delta I_t$. For a clear day the removal of the deterministic variability yields very small fluctuations of $\Delta I_t$ whereas for the cloudy days the large ramps manifest themselves in the $\Delta I_t$ curve.

### 2.6.2. Forecast uncertainty

Uncertainty of the forecast, as defined in Ref. [27], is a normalized version of the RMSE used in Ref. [42], which is similar to the
relative RMSE used in Ref. [43,44]. The difference here is that normalization is performed with respect to \( I_{dir} \), rather than the average irradiance, as

\[
U = \sqrt{\frac{1}{N} \sum_{t=1}^{N} \left( \frac{I_t - I_{dir,t}}{I_{dir,t}} \right)^2}.
\]

(25)

This definition of forecast uncertainty is the second required quantity for the determination of the performance metric described in the next section.

### 2.6.3. Performance metric

Now that solar variability \( V \) and forecast uncertainty \( U \) have been defined, we can write the performance metric suggested in Ref. [27] as the difference between \( V \) and \( U \) normalized with respect to \( V \)

\[
s = \frac{V - U}{V} = 1 - \frac{U}{V},
\]

(26)

where it is implied that \( U \) and \( V \) are calculated using the same data set. The authors in Ref. [27] point out that when \( s = 1 \) the forecast performs perfectly due to the vanishing uncertainty, however, when \( s = 0 \) the forecast is dominated by solar variability. Close inspection of Equations (10), (24) and (25) reveals that the persistence forecast, by definition, should have a performance metric value of \( s = 0 \) (see Table 1). As a result, the ratio \( U/V \) can be thought of as a measure of the quality of performance of a forecast with the persistence model as the baseline. If a developed forecast model produces a performance metric for which \( s < 0 \), the model performs worse than a persistence forecast. This means that a typical forecast model should be characterized with values ranging from 0 to 1, with values closer to 1 indicating a higher quality forecast [27].

As a result of \( U \) and \( V \) being random variables, \( s \) is also a random variable. Thus, the authors in Ref. [27] suggest using the average value of the metric (\( s \)) for a representation of forecasting skill. In order to do this, the data is partitioned into several time windows of equal length \( N_w \). Values of \( U_i \) and \( V_i \) are calculated for each of the \( j \) windows and an average value of the metric (\( s \)) is obtained. An important feature of the performance metric is that periods of low variability do not positively bias the value of (\( s \)). That is to say, when a time window has many clear days (low variability) the forecasting error will also be low, therefore preserving the relative amount of \( U \) and \( V \), as seen in Fig. 4. The opposite is also true. Events of high variability which are completely predictable, such as an eclipse, will not negatively bias the value of the performance metric because uncertainty is near zero.

Marquez and Coimbra applied the preceding analysis to two forecasting models based on Artificial Neural Networks (NAR and NARX) for hourly data collected from Jan. 1–Oct. 31 of 2010, and found that the metric is a statistical invariant which is preserved over a wide variety of time horizons [27]. This makes the metric an attractive and robust candidate for the comparison of solar forecast model performance. Fig. 4 shows scatter plots of \( U_j \) versus \( V_j \) computed for each \( j \)th time window for which \( N_w = 50, 100, 150 \) and 200, and Table 1 gives corresponding numerical values of \( s \) obtained using \( N_w = 200 \). The ESRA clear sky model was employed for normalization purposes. General trends in the scatter plots demonstrate that the performance of the NAR forecast model is similar to the persistence forecast’s. On the other hand, the NARX model shows significant forecasting improvement over the persistence as many of the points fall below the 1:1 reference line.

### 3. Regressive methods

Prior to stochastic techniques solar radiation models focused on the extrapolation of data from long term averages and steady state values resulting in essentially static models which described only seasonal and diurnal changes. Landsberg [45] incorporated GHI data from over 300 locations into a world map of average solar energy available at the Earth’s surface. Whillier [46] derived graphical models from correlations of measured GHI in several parts of the world at latitudes ranging from the equator to 50° north and south. Swartman and Ogunlade [47,48] derived relationships correlating solar radiation intensity with sunshine and relative humidity for tropical conditions in the Toronto area. Lund [49] correlated nine years of daily radiation data taken at Blue Hill, Massachusetts, month by month with observations of snow cover, wind, sunshine, sky cover, pressure and precipitation in order to find a linear combination of parameters to best estimate GHI. These models, whether graphical or mathematical, ignored the short-term time dependent patterns of solar radiation data which result from changing local weather conditions and cause fluctuations on timescales ranging from seconds to days. Forecasting of such fluctuations is essential to the operation of, for example, concentrating photovoltaic systems due to their relatively small apertures and strong dependence on DNI [50]. Any forecasting model which ignores the short term stochastic characteristics of solar radiation data is clearly deficient.

A second shortcoming of the pre-stochastic models was the use of common regression techniques to develop mathematical relationships. The statistical theory at the foundation of these regression techniques assumes that the individual observations of solar irradiance change independently [51]. However the reverse is true and changes due to local weather conditions generate data time-series which are strongly dependent. The correlated nature of solar irradiance data is another essential characteristic which must be addressed by any successful forecasting model. It is evident from the discussion that models which are based on long term averages and employ independent regression techniques could be improved through the use of stochastic models which account for short term fluctuations and the correlated nature of the data.

In the 1920s G. Udny Yule began to establish new approaches for the analysis of stationary time-series. Yule was prompted by the quintessential criticism of strict periods which is at the heart of the hypothesis of the classical methods of Fourier and Schuster [52]. While these methods are pervasive in almost all areas of the physical sciences, the rigid nature of strict periods limits the applications for which these models are adequate. In response to the call for a modified method, in 1922 Yule pioneered the first of two new approaches for stationary time-series analysis which he called Moving Averages (MA), see Fig. S(c). The theory originates from two papers regarding the regular fluctuations of purely random series which are given, for instance, by throwing dice [53,54]. The second of Yule’s approaches, linear Auto-Regression (AR), Fig. 5(b), came in 1927 in a well known paper which discussed the periodicities of disturbed series with special reference to Wolfer’s sunspot

### Table 1

<table>
<thead>
<tr>
<th>Model</th>
<th>( R^2 )</th>
<th>RMSE (W/m²)</th>
<th>( s )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ESRA</td>
<td>0.969</td>
<td>55.5</td>
<td>0</td>
</tr>
<tr>
<td>NAR</td>
<td>0.972</td>
<td>53.2</td>
<td>2.27%</td>
</tr>
<tr>
<td>NARX</td>
<td>0.977</td>
<td>48.8</td>
<td>12.02%</td>
</tr>
<tr>
<td>Validation set</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ESRA</td>
<td>0.926</td>
<td>59.5</td>
<td>0</td>
</tr>
<tr>
<td>NAR</td>
<td>0.924</td>
<td>60.2</td>
<td>-2.53%</td>
</tr>
<tr>
<td>NARX</td>
<td>0.949</td>
<td>49.4</td>
<td>16.25%</td>
</tr>
</tbody>
</table>
numbers [55]. Shortly after, in the 1930s, the work of Yule was furthered by H. Wold who applied the theory to business cycles and econometric statistics [52]. It wasn’t until the 1960s that G. E. P. Box and G. M. Jenkins, motivated by the principle of parsimony, would popularize what would become two of the most widely used models in time-series analysis [56]. The first, known as Auto-Regressive Moving Averages (ARMA), advanced from the combination of the schemes established by Yule and Wold for stationary time-series analysis, Fig. 7(a). The second, utilized an alternative component and gave way to a class of models known as Auto-Regressive Integrated Moving Averages (ARIMA) which are useful for the analysis of non-stationary processes, Fig. 7(b) [52]. The use of ARMA and ARIMA provide a basis for many problems outside the realm of solar irradiance forecasting including economic and business planning, production planning, inventory and production control and optimization of industrial processes [56].

Before beginning a discussion of the various stochastic models it will be useful to review some simple operators and terminology. One operator which will be used frequently is the forward shift or delay operator \( q \) which, when applied to a time-series \( \{z_t\} \) at time \( t \), is defined as

\[
qz_t = z_{t+1}.
\]  

(27)

The forward shift operator can be applied successively to yield

\[
q^k z_t = z_{t+k}.
\]  

(28)

The inverse of the forward shift operator is the backward shift or delay operator \( q^{-1} \) which, when applied to a time-series, is given by

\[
q^{-1} z_t = z_{t-1}.
\]  

(29)

which in turn yields,

\[
q^{-k} z_t = z_{t-k}.
\]  

(30)

The backward shift operator can be used to construct the backward difference operator \( D \) as follows,

\[
Dz_t = z_t - z_{t-1} = (1 - q^{-1}) z_t.
\]  

(31)

Its inverse, the summation operator \( S \), can be written as

\[
Sz_t = D^{-1} z_t = z_t + z_{t-1} = (1 + q^{-1}) z_t.
\]  

(32)

The stochastic models that follow are based on the result from Yule [55] that if a strong dependency exists between successive terms in a time-series which is, for example, a characteristic of solar irradiance data, it can be effectively generated by a series of independent shocks which are assumed to have non-permanent effects and are drawn at random from a stationary distribution having zero mean and variance \( \sigma_u^2 \). Any series of such random shocks \( \{u_t\} \) is
stationary summation ARIMA process is said to be regressed on the previous values. If we de  

The level of a stationary process can be thought of as the average  

to section, as exempli  

non-stationary process is time dependent and varies from section  

Fig. 7.  

stationary and non-stationary. If the sequence of weights in Equa-  

rations in many areas of the physical sciences. For instance, obser-  

value of the process can be expressed as a  

of the previous values of the process and a single shock  

uctuates. However, the level of a  

uctuates as seen in Fig. 6(a). Stationary time-series  

onomenon with time can be classi  

ments;  

fluctuates, or in  

finite, or in  

finite which gives us an inter-  

ed by the following operation  

arbitrary process to be stationary a set of condi-  

is a parameter which determines the level of the process. The level of a stationary process can be thought of as the average value about which the series fluctuates. However, the level of a non-stationary process is time dependent and varies from section to section, as exemplified in Fig. 6. The operator that transforms  

where  

m = 1. It follows that the AR(m) model may be written conveniently as  

Here it is clear that the process is regressed on the previous values of  

Therefore, it is helpful to think of the AR(m) process as the output of a linear filter with transfer function  

which allows expansion of the process in partial fractions,  

3.1. Linear stationary models  

Observational series that describe a changing physical phenomenon with time can be classified into two main categories; stationary and non-stationary. If the sequence of weights in Equation (34) is finite, or infinite and convergent, the linear filter is said to be stable and the time-series  

stationary time-series are static with respect to their general shape. The fluctuations may appear, to the user, as clearly ordered or totally random, nonetheless, the character of the series is, on average, quite similar in different segments. In this case, the parameter  

be interpreted as the average value about which the series fluctuates as seen in Fig. 6(a). Stationary time-series find applications in many areas of the physical sciences. For instance, observational time-series and series involving deviations from a trend are often stationary [52]. In fact, the stochastic portion a solar radiation data set is often framed as a stationary process [57].  

3.1.1. Auto-Regressive (AR) models  

The AR models get their name from the fact that the current value of the process can be expressed as a finite, linear combination of the previous values of the process and a single shock . Thus, the process is said to be regressed on the previous values. If we define the stochastic portion of the time-series  

then the Auto-Regressive process of order  

We can simplify the previous expression by defining the Auto-Regressive transfer function of order ,  

where  

where it is clear that if  

3.1.2. Moving Average (MA) models  

While the AR techniques model the stochastic portion of the time-series  

the term moving average can be a bit misleading due to the fact that the weights in Equation (42) do not, in general, need to be positive nor does their sum necessarily equal unity [56]. Nonetheless, the name is used for historic convention. The MA(n) operator is defined as  

Let us pause here and note that the term moving average can be  

(35)  

(36)  

(37)  

(38)  

(39)  

(40)  

(41)  

(42)

Fig. 7. (a) ARMA(n,m) modeled as a linear filter with transfer function and (b) ARMA(m,n) modeled as a stationary ARMA(m,n) linear filter in series with a non-stationary summation filter  

(33)  

(34)
\[ \Theta_n(q) = \sum_{k=0}^{n} \theta_k q^{-k} = 1 + \theta_1 q^{-1} + \theta_2 q^{-2} + \ldots + \theta_n q^{-n}, \]  
(43)

where \( \theta_0 = 1 \). As a result we can write the MA model in an economic fashion

\[ z_t = \Theta_n(q) \omega_{t}. \]  
(44)

The MA process can be thought of as the output \( z_t \) of a linear filter whose transfer function is \( \Theta_n(q) \), with white noise \( \omega_{t} \) as the input (see Fig. 5(c)). Like its counterpart, the MA model contains \( n \) undetermined parameters \( \theta_1, \ldots, \theta_n, \) and the variance of the white noise process \( \sigma^2_{\omega} \) which must be determined from the data using the techniques described in the next section. Unlike AR models, MA models are unconditionally stable [52].

3.1.3. Mixed Auto-Regressive Moving Average (ARMA) models

Linear processes represented by an infinite or an extraneous number of parameters are clearly not practical. However, it is possible to introduce parsimony and still obtain useful models. A well known result in time-series analysis is the relationship between the \( \Theta \) weights and the \( \Phi \) weights [56]. Operating on both sides of Equation (38) by \( \Theta(q) \) and making use of Equation (44), yields

\[ \Theta(q) \Phi(q) z_t = \Theta(q) \omega_t = \hat{z}_t \]  
(45)

which implies

\[ \Theta(q) \Phi(q) = 1 \]  
(46)

that is

\[ \Phi^{-1}(q) = \Theta(q). \]  
(47)

Equation (47) indicates that the \( \Phi \) weights may be arrived at from knowledge of the \( \Theta \) weights, and vice-versa. Thus the finite MA(n) process \( \hat{z}_t = \Theta_n(q) \omega_t \) can be written as an infinite AR process

\[ \hat{z}_t = \omega_t - \theta_1 \hat{z}_{t-1} - \theta_2 \hat{z}_{t-2} - \ldots \]  
(48)

However, if the process were really MA(n), we would arrive at a non-parsimonious representation in terms of the AR expression. By the same reasoning, a finite AR(m) model could not be parsimoniously represented using an infinite MA expression. Therefore, in practice, in order to realize a parametrization which is parsimonious, both AR and MA terms are often used in the model development. Hence,

\[ \hat{z}_t + \phi_1 \hat{z}_{t-1} + \ldots + \phi_m \hat{z}_{t-m} = \omega_t - \theta_1 \omega_{t-1} - \ldots - \theta_n \omega_{t-n} \]  
(49)

or

\[ \Phi_m(q) \hat{z}_t = \Theta_n(q) \omega_t. \]  
(50)

Equation (50) is referred to as the mixed Auto-Regressive Moving Average (ARMA) process of order \((m, n)\). It is illustrative to note that the ARMA\((m, n)\) model can be written as

\[ \hat{z}_t = \Theta_n(q) \omega_t / \Phi_m(q) = 1 + \theta_1 q^{-1} + \ldots + \theta_n q^{-n} / 1 + \phi_1 q^{-1} + \ldots + \phi_m q^{-m} \omega_t \]  
(51)

where it is clear that \( \hat{z}_t \) is the output from a linear filter, whose transfer function is the ratio of two polynomials \( \Theta_n(q) \) and \( \Phi_m(q) \), with white noise \( \omega_t \) as the input (see Fig. 7(a)).

In practice, it is frequently true that adequate representation of actually occurring stationary time-series can be obtained from models in which \( n \) and \( m \) are not greater than two and often less than two [56,57]. The order of the model, that is the values of \( m \) and \( n \), is determined using the sample auto-correlation function and partial auto-correlation function of the time-series [59]. The model parameters are estimated by least squares methods and the resulting model is said to adequately describe the statistical information contained in the series in a parsimonious manner.

These techniques have been used since the 1970s to analyze time-series resulting from solar irradiance data. Boileau [60] and Guerrier [61] evaluated several stochastic models in the field of solar meteorology for 24-h time steps. Later, Guerrier, Boileau and Bernard [62] would employ seasonal and non-seasonal stochastic techniques for solar irradiance prediction which were originally developed by Box and Jenkins. Guerrier [63] showed the usefulness of small correlations of consecutive day solar irradiance data through an optimization of stochastic techniques. Brinkworth [64] used the simple ARMA processes to model irradiance in cloudy areas of the U.K. in order to predict the output of solar thermal systems. Benard et al. [65] estimated empirical values for stochastic models from irradiance measurements performed at three sites: a temperate site (Trappes, France, 49N), a Mediterranean site (Carpentras, France, 44) and an equatorial altitude site (Huallao, Peru, 13S, 3 km). Subsequently, these values were used to develop ARMA models for the prediction of daily solar irradiance [65]. Mungan et al. [67] compared ARMA models with factor analysis, Markov transition-matrices, Gaussian mapping and transmittance transition tensor techniques. Boh et al. [68] attempted to predict solar irradiance at hourly time steps using stochastic techniques for locations in France. ARMA processes have also been used to model hourly irradiance for the calculation of the optimal control of buildings in France and Japan [69–71]. Aguiar and Collares-Pereira modeled daily sequences of hourly irradiance and clearness index \( K_r \) using ARMA techniques, where \( K_r \) was obtained by multiplying a clear sky value by a non-stationary fluctuation, with probability depending on the hour of the day [72]. Later, Aguiar and Collares-Pereira used an ARMA(1, 0) model to generate hourly series of solar irradiance values which were transformed backwards to generate synthetic sequences of clearness index \( K_r \) values [73]. Al-Awahdi et al. proposed an ARMA model which used a bilinear time-series to generate hourly irradiance models for Kuwait [74]. Mora-Lopez et al. [75] employed a multiplicative ARMA model for global radiation time-series with diurnal and seasonal components. The multiplicative nature of these models enabled the acquisition of two types of relationships observed in recorded hourly series of global irradiation: the relationship between the value at 1 h and the value at the previous hour, and the relationship between the value at 1 h in one day and the value at the same hour in the previous day. Moreno-Muñoz et al. [76] more recently used multiplicative ARMA models to generate instantaneous series of global irradiance in southern Spain during a four year period. Lately, however, ARMA methods typically find applications as components of a robust hybrid system architecture [77–81] as explained in Section 8.

3.1.4. Mixed Auto-Regressive Moving Average Models with exogenous variables (ARMAX)

All of the linear stationary stochastic techniques discussed so far have been univariate, meaning the technique uses previous values of only the time-series it is attempting to model. However, the accuracy of ARMA\((m, n)\) models may be improved by including information external to the time-series under analysis. For example, in the case of solar irradiance forecasting, the error of a forecasting model may be reduced by including information about the evolution of the local temperature, relative humidity, cloud cover, wind speed, wind direction, etc. Variables such as these, which are independent of the model but affect its value, are referred to as exogenous variables. We can include into the
ARMA(m, n) models p exogenous input terms which allow us to write the ARMAX(m, n, p) process as

\[ \ddot{z}_t + \phi_1\ddot{z}_{t-1} + \cdots + \phi_m\ddot{z}_{t-m} = \omega_t + \theta_1\omega_{t-1} + \cdots + \theta_n\omega_{t-n} + \lambda_1e_{t-1} + \cdots + \lambda_pe_{t-p}. \] (52)

The above model contains AR(m) and MA(n) models as well as the last p values of an exogenous time-series \( e_t \). Defining the exogenous input operator of order \( p \) as

\[ \Lambda_p(q) = \sum_{k=1}^{p} \lambda_k q^{-k} = \lambda_1 q^{-1} + \lambda_2 q^{-2} + \cdots + \lambda_p q^{-p} \] (53)

allows us to write the ARMAX(m, n, p) model conveniently as

\[ \Phi(q)\ddot{z}_t = \Theta_n(q)e_t + \Lambda_p(q)e_t. \] (54)

The careful reader might already be aware of the fact that all of the linear stationary models discussed so far have a similar structure. In fact, many models in linear system analysis can be considered a special case of the general discrete time model structure

\[ \Phi(q)\ddot{z}_t = \Theta(q)e_t + \Lambda(q)e_t \] (55)

where \( \Phi(q), \Theta(q), \Lambda(q), \Psi(q) \) and \( \Xi(q) \) are polynomials of the shift operator \( q \). Table 2 summarizes some of the commonly used discrete time models which can be considered special cases of (55), however, the current study is limited to the models we have discussed so far.

### 3.2. Non-linear stationary models

So far we have only considered general classes of linear stationary models. However, non-linear methods would enable powerful structures with the ability to accurately describe complex non-linear behavior such as chaos, hysteresis and saturation effects or a combination of several non-linear problems [82]. A step towards non-linear modeling is made by introducing the Non-linear Auto-regressive eXogenous (NARX) model as

\[ \ddot{z}_t = f(\ddot{z}_{t-1}, \ddot{z}_{t-2}, \ldots, \ddot{z}_{t-m}, e_{t-1}, e_{t-2}, \ldots, e_{t-n}) + \omega_t. \] (56)

In much the same way one can also convert the ARMAX model into a Non-linear ARMAX model (NARMAX) as follows

\[ \ddot{z}_t = f(\ddot{z}_{t-1}, \ddot{z}_{t-2}, \ldots, \ddot{z}_{t-m}, e_{t-1}, e_{t-2}, \ldots, e_{t-n}, \omega_{t-1}, \omega_{t-2}, \ldots, \omega_{t-p}) + \omega_t. \] (57)

These non-linear input–output models find many applications in the field of engineering, especially in the parametrization of Artificial Networks which are discussed in Section 4.1.

### 3.3. Linear non-stationary models

If the sequence of weights in Equation (34) is infinite but not convergent, the linear filter's transfer function \( G(q) \) is said to be unstable and the process \( z_t \) to be non-stationary. In this case, \( \mu \) has no physical meaning except as a reference to the level of the process (Fig. 6(b)). Non-stationary processes are different in one or more respects throughout the time-series due to the time dependent nature of the level. As a result, in the analysis of non-stationary time-series, time must play a fundamental role, for example, as the independent variable in a progression function, or as a normalization factor in the analysis of the evolution of a phenomenon from an initial state [52]. Several observed time-series behave as if they have no specified mean about which they fluctuate, for example, daily stock prices or hourly readings from a chemical process [56].

### 3.3.1. Auto-Regressive Integrated Moving Average (ARIMA) models

While non-stationary processes do not fluctuate about a static mean, they still display some level of homogeneity to the extent that, besides a difference in local level or trend, different sections of the time-series behave in a quite similar way. These non-stationary processes may be modeled by particularizing an appropriate difference, for example, the value of the level or slope, as stationary. (Fig. 6(b)(c)). What follows is a description of an important class of models for which it is assumed that the dth difference of the time-series is a stationary ARMA(m, n) process.

We have seen that the stationarity condition of an ARMA(m, n) process is that all roots of \( \Phi_m(q) = 0 \) lie outside the unit circle, and when the roots lie inside the unit circle, the model exhibits non-stationary behavior. However, we have not discussed the situation for which the roots of \( \Phi_m(q) = 0 \) lie on the unit circle. Let us examine the following ARMA(m, n) model

\[ \Phi_m(q)\ddot{z}_t = \Theta_n(q)e_t \] (58)

and specify that \( d \) of the roots of \( \Phi_m(q) = 0 \) lie on the unit circle and the residue lie outside. We can then express the model as

\[ \Phi_m(q)\ddot{z}_t = \Theta_n(q)(1 - q^{-1})^d \ddot{z}_t = \Theta_n(q)e_t \] (59)

where \( \Phi_m(q) \) is a stationary and invertible AR(m) operator. Seeing that \( S^d\ddot{z}_t = S^d\ddot{z}_t \) when \( d \geq 1 \), we can write

\[ \Phi_m(q)S^d\ddot{z}_t = \Theta_n(q)e_t. \] (60)

Defining \( y_t = S^d\ddot{z}_t \) allows one to express the model in a more illustrative way

\[ \Phi_m(q)y_t = \Theta_n(q)e_t \] (61)

where it is clear that the model is in agreement with the assumption that the dth difference of the time-series can be regarded as a stationary ARMA(p, q) process. We refer to (60) as the Auto-Regressive Integrated Moving Average (ARIMA) process. Because the AR transfer function \( \Phi_m(q) \) is of order \( m \), \( d \) differencing steps are taken, and the MA transfer function \( \Theta_n(q) \) is of order \( n \) in (60), we refer to the process as ARIMA(m, d, n). In practice, \( d \) is typically 0, 1 or at most 2 [56]. As mentioned above, the ARIMA(m, d, n) model is equivalent to representing the process \( z_t \) as the output of a linear filter with transfer function \( \Phi_m^{-1}S^d\Theta_n \) and takes white noise \( \omega_t \) as an input (Fig. 7).

In the past ARIMA models have found applications in many areas of research, however, only recently have they been employed to model problems involving solar radiation. Craggs et al. [83] used
ARIMA models to compare 10 min, 20 min, 30 min and 1 h averages of solar irradiance levels for a site in the UK. Santos et al. [84] used an ARIMA(1, 1, 1) model to calculate synthetic daily solar radiation values using the monthly average radiation as the input for locations in Spain. Kärner [85] carried out an ARIMA analysis for satellite-based global tropospheric and stratospheric temperature anomalies and solar irradiance data sets. Kärner’s results emphasized a dominating role of the solar irradiance variability in variations of the tropospheric temperature. Yüreklı et al. [86] imposed ARIMA based techniques to generate solar irradiance, temperature and relative humidity forecasts for use in the control strategy of agricultural facilities. Stanhill and Cohen [87] used a first order ARIMA model to describe the time course of annual sunshine duration and global irradiance to study trends and changes in solar forcing at the Earth’s surface and reported a 0.5% increase in solar forcing per decade during the 20th century. Kärner [88] later employed ARIMA models to compare the variability of the total solar irradiance at the top of the atmosphere to surface air temperature series, which he reports are strongly correlated. Reikard [89] compared ARIMA models with a number of other methods, including transfer functions, neural networks and hybrid models, for six data sets at resolutions of 5, 15, 30 and 60 min using the global horizontal component of solar irradiance. Reikard found that, in nearly all the tests, the best results are obtained using the ARIMA in logs, with time-varying coefficients. In Ref. [90] the authors used daily solar radiation measurements carried out in Bogotá, Columbia, from 2003 to 2009 as inputs for an ARIMA(1, 0, 0) model for forecasting the mean daily global solar radiation. Like its stationary counterpart, the ARIMA model has, as of late, also found many applications in the construction of hybrid systems [79–81,89].

3.3.2. Auto-Regressive Integrated Moving Average models with exogenous variables (ARIMAX)

In a similar way to the ARMAX(m, d, n, p) model, the previous p values of an exogenous time-series εt may also be included into the ARIMA(m, d, n) model to yield the ARIMAX process of order (m, d, n, p)

\[ z_t = \phi_1 z_{t-1} + \ldots + \phi_m z_{t-m} + \omega_t + \theta_1 \omega_{t-1} + \ldots + \theta_n \omega_{t-n} + \lambda_1 e_{t-1} + \ldots + \lambda_p e_{t-p} \]  

(62)

As we did before, defining \( y_t = S_d z_t \) in terms of the backwards shift operator allows us to express the model in a more compact form

\[ \Phi_m(q) y_t = \Theta_n(q) \omega_t + \Lambda_p(q) e_t \]  

(63)

which again looks very similar to Equation (55).

4. Artificial Intelligence (AI) techniques

The development of Artificial Intelligence (AI) techniques began in the early 1950s with a number of experiments conducted by Herbert Simon, Allen Newell and Cliff Shaw [91]. Simon was consulting at the RAND Corporation when he saw a printer using an anomaly and solar irradiance data sets. Meanwhile, Claude Shannon demonstrated how a machine could be used to play a reasonable game of chess [94]. Shannon’s chess program relied on the optimization of a weighted function dependent on the position of the chess pieces using a minmax procedure. Admitting that the work of Simon et al. and Shannon was the earliest in the field artificial intelligence, the topic’s origin is frequently associated with another event.

The first coordinated undertaking in the field of machine intelligence was the Artificial Intelligence Conference organized by John McCarthy, Marvin Minsky, Nathaniel Rochester and Claude Shannon at Dartmouth College in New Hampshire in 1956 [91]. It was at this conference that John McCarthy, who would develop the popular LISP family of computer programming languages two years later, suggested the term “Artificial Intelligence”, which is still frequently used today to describe a wide variety of machine learning methods. The AI Conference of 1956 is commonly viewed as the event responsible for the motivation to examine computers to solve problems rather than focusing on hardware that simulated intelligence [95].

Several AI methods were developed in the years following the conference in New Hampshire. In 1959 Simon et al. developed a General Problem Solver (GPS) program capable of proving mathematical theorems like Logic Theorist, playing chess reasonably like Shannon’s algorithm and, in addition, solving elaborate geometric puzzles. GPS accomplished this by generating heuristics through means-ends analysis [96]. However, the predominant criticism of the GPS code was that it could not learn from problems solved in advance [91]. This paved the way for McCarthy’s LISP programming language to become the standard in AI development [97].

In 1960 additional codes were developed that mimicked key aspects of human thought, such as language and vision processing. Kenneth Colby, an American psychiatrist working at Stanford University, developed a family of computer programs known as “chatbots” which were able to simulate intelligent conversation with people through a process known as Natural Language Processing (NLP). Soon after, Joseph Weizenbaum wrote a program known as ELIZA which would become one of the most popular chatbots [91]. ELIZA used rules correlated with words like ‘I’, ‘you’, ‘like’ etc., which were prompted if one of these words was encountered [98]. Bobrow [99] and Winograd [100] also made contributions to the emerging field of NLP, most notably through the development of programs with the ability to make assumptions and to learn from previously solved problems [91]. Parallel to this work, Marvin Minsky and his group at MIT, which supported the idea that AI is fundamentally a symbol processing exercise [101], developed a computer program that could perform visual analogies by identifying visual relationships from two figures and locate another set of figures with similar relationships from a larger set [102].

All of the methods mentioned above use a “symbolic” approach to intelligent systems. This symbol processing architecture focuses on knowledge representation through the translation of the user’s expert knowledge into a set of formal symbols. Once a formal encoding is established the system functions in the traditional Von Neumann computer pattern: INPUT → PROCESS → OUTPUT. The advantage lies in the enormous representational power of these systems which stems from their generality. However, the central criticism is that only the user’s own knowledge comes into effect [103]. In addition, these systems are limited by the fact that all the responses must be explicitly programmed beforehand. These symbolic approaches are referred to as brittle due to the “crashing” that occurs when an unusual set of circumstances not anticipated in the response coding occurs. Even so, symbol processing remained the popular motif in AI systems development well into the 1980s.
However, there was another school of thought that defended that intelligent systems should be modeled after living organisms with brains and the ability to adapt.

A seminar paper in the Bulletin of Mathematical Biology in 1943 titled “A Logical Calculus of Ideas Immanent in Nervous Activity” [104] was the result of a joint effort between Warren McCulloch and Walter Pitts. In this work, which would later be included in Hebb’s famous Organization of Behavior [105], the authors proposed the first artificial neuron called the Threshold Logic Unit. In 1958 Frank Rosenblatt, a computer scientist working at Cornell Aeronautical Laboratories, published a book titled “Principles of Neurodynamics” [106]. In this book, Rosenblatt discussed machines with the ability to learn how to classify information by adapting weights, which he called perceptrons. The Pandemonium architecture, which was proposed by Oliver Selfridge in 1959 [107] and saw success in modeling human pattern recognition, suggested that the brain works like a collection of daemons, or background processes, which work in parallel and are each responsible for a single task. Soon after, at Stanford University, Professor Bernard Widrow and his graduate student Ted Hoff developed the ADALINE (ADAptive Linear NEuron) and the Least Means Square (LMS) rule for training [108]. This work resulted in the emergence of a new discipline of AI techniques known as Artificial Neural Networks (ANNs). ANNs are essentially attempts to create an electrical analog to the biological neural networks present in the human brain.

Another set of AI techniques that were motivated by a living organism’s ability to adapt and evolve was developed in the 1970s by John Holland, a professor of Psychology and Electrical Engineering at the University of Michigan, Ann Arbor. In his book titled “Adaptation in Natural and Artificial Systems” [109], Holland, motivated by Darwinian views of evolution, conceptualized an algorithm which would identify an optimal solution through the combination and reproduction of the strongest individuals in a solution space. Holland’s work was the first in a branch of AI techniques known as Genetic Algorithms (GAs). Neural and genetic computing remained relatively unpopular during the 1970s, however, with an increase in available computing power and technological advancements such as multi-layer perceptrons, the 1980s saw a renewed interest in ANNs and GAs [95,110].

In contrast to the symbolic approach, these “evolutionary” approaches emphasize learning and adaptation over the representational power of symbolic processing. The “intelligence” of these systems is attributed to the ability to distinguish between two general classes of patterns without being directly programmed [103]. For example, a set of solar irradiance measurements in one location forms a class of patterns which is different from a set at another location, as long as these sites are far enough away from each other as to not be correlated through local weather patterns [40]. Evolutionary techniques are much more systematic than symbolic approaches and it should be noted that the internal structure of the system itself is critical inasmuch as it determines the type of function the ANN is able to approximate [103].

So far, we have discussed several AI techniques. It is beneficial to pause here and take note of the several branches into which these techniques can be classified.

- **Problem Solving and Planning**: systematic refinement of goal hierarchy and plan revision mechanisms, e.g. Shannon’s chess code.
- **Expert Systems**: uses knowledge-base of previously established rules of thumb for well-defined problems for complex decision-making, e.g. Simon’s Logic Theorist and GPS
- **Natural Language Processing**: text processing and generation, speech synthesis and analysis, and machine translation, e.g. Weizenbaum’s ELIZA
- **Computer Vision**: Image understanding, facial recognition, motion derivation and intelligent visualization; e.g. Minsky’s visual analogy program.
- **Genetic Algorithms**: Evolutionary algorithms with an inherent learning capacity.
- **Artificial Neural Networks**: Combination of pattern recognition, deductive reasoning and numerical computations to simulate learning in the human brain.
- **Hybrid Systems**: Any number of combinations of the above branches.

The present study, however, focuses on ANNs, GAs and their applications in various Hybrid Systems (HS). For a comprehensive review of AI techniques see Refs. [91,95]. Before moving on, it should be noted that while there have been several attempts to define AI [111–114], no single definition has been universally accepted. The authors in Ref. [95] suggest that this is likely due to the various definitions of “intelligence”, which is an abstract and immeasurable quantity.

### 4.1. Artificial Neural Networks (ANNs)

ANNs were motivated from the observations of a special type of cell known as a neuron. Neurons are responsible for a number of signal processing tasks in our bodies such as responding to touch, sound, light or other external stimuli and receiving signals from the brain and spinal chord which control muscles, glands and arteries. Neurons are located in the brain, spinal chord and the autonomic ganglion found just outside of the spinal chord. Neurons are furnished with a number of antenna-like structures that stretch out from the cell body, or soma, which allow the cell to send and receive signals from other cells and the environment. The structures which allow the neuron to accept input signals are called dendrites, while the structures which carry signals away from the neuron are called axons. A neuron may posses numerous dendrites but it never has more than one axon. Nonetheless, the dendrites and axon may branch hundreds of times before they terminate forming complex tree-like structures as seen in Fig. 8(a). By actively regulating calcium, chloride, potassium and sodium ion concentrations inside the cell, the neuron is able to maintain electrical potential gradients across its membrane. When a neuron’s electric potential exceeds a specific threshold, an all-or-none electro-chemical impulse called an

---

**Fig. 8.** (a) Simplified diagram of a biological neuron. Antenna-like structures which extend from the cell body, or soma, allow the neuron to communicate with other cells. The structures which allow the neuron to accept input signals are called dendrites. The structure which carries signals away from the neuron is called an axon. A neuron may posses numerous dendrites but it never has more than one axon. (b) Artificial neuron with inputs \( x_1, x_2 \) and \( x_3 \) weighted by \( w_1, w_2 \) and \( w_3 \). The neuron has an embedded net function \( \beta = \sum_{j=1}^{3} w_j x_j \) and transfer function \( f(\cdot) \) which are used to calculate output \( z \).
action potential is generated. The action potential speeds down the neuron’s axon and triggers synaptic communications at the tips of the axonic tree. This process is known as the “firing” of a neuron. The rate at which the neuron fires, which can reach up to 300 Hz, is regulated by adrenaline which acts as a bias for the neuron, making it much more likely to fire in the presence of a stimulus [115].

Once the architecture of a biological neuron is understood, one can begin to construct an artificial neuron. Similar to their biological counterparts, artificial neurons can be connected to each other to form a network. A simple model of an artificial neuron which takes inputs \( x_1, x_2 \) and \( x_3 \) from three other neurons and processes a corresponding output \( z \) is detailed in Fig. 8(b). The quantity \( \beta \) is known as the bias and is used to model adrenaline’s ability to lower the threshold. Because not every input is equally relevant to whether the neuron fires or not, we assign weights \( w_1, w_2 \) and \( w_3 \) to each of the inputs. Here we use a linear net function to sum the inputs which is given by

\[
y = \beta + \sum_{i=1}^{3} w_i x_i, \tag{64}
\]

The firing rate of the action potential is modeled through the use of a transfer function \( f(\cdot) \). Several different transfer functions have been suggested over the years (see Table 3). The most commonly used transfer function is the sigmoid or logistic function due to its attractive mathematical properties such as monotonicity, continuity and differentiability [95,115]. Employing the sigmoid transfer function with the linear net function, the output \( z \) of the artificial neuron in Fig. 8(b) can be written in a very simple manner

\[
z = \frac{1}{1 + e^{-(\beta + \sum_{i} w_i x_i)}} \tag{65}
\]

The attractiveness of ANNs stem directly from their likeness to their biological counterpart. Specifically, ANNs allow for very low level programming (net and transfer functions) to solve a wide variety of complex, non-linear, non-analytic, non-stationary, stochastic or general mathematically ill-defined problems in a self-organizing manner that requires little or no interference with the program itself. Conventional computing algorithms apply elaborate sets of equations to solve specific and well defined problems through the programmer’s imposed organization. Clearly, this is not the most parsimonious expression of a solution. When a pesky fly avoids a dedicated flyswatter, it has no time to carefully solve complex differential equations concerning Newton’s laws of motion. Nonetheless, its non-specific decision making process allows it to consistently stay one step ahead of the flyswatter, in addition to carrying out a wide number of other tasks. In order to be as efficient and versatile as possible biological systems must converge to the simplest algorithmic architecture [116].

Table 4 summarizes the early progress in ANN structure development which is discussed in the next section. Careful examination of Table 4 reveals a twenty-two year dormant period in ANN development between 1960 and 1982. This hibernation resulted from the fact that networks with hidden layers were not considered. These single layer ANNs were unable to solve the popular two state eXclusive-Or (XOR) problem and researchers lost interest as a result [82]. Despite that, the 1980s saw an explosion in renewed interest in ANNs with the development of multi-layer perceptrons and back propagation algorithms. Thus, as one can see, the inner structure of ANNs is particularly relevant. The inner-structures of ANNs are typically described with what is commonly known as a directed graph. Directed graphs consist of neurons and directed arcs which describe the synaptic links. The following subsections cover the most widely used ANN structures in more detail.

4.2. Early networks

4.2.1. Threshold Logic Unit (TLU)

The earliest formulation of an artificial neuron was the Threshold Logic Unit (TLU) developed by McCulloch and Pitts in 1943. The TLU employed, as a transfer function, the threshold or Heaviside step function (see Table 3). In Ref. [104] the authors stated five assumptions which governed their neuron:

1. The activity of a neuron is an “all-or-none” process.
2. A certain fixed number of synapses must be excited within the period of latent addition in order to excite a neuron at any time, and this number is independent of previous activity and position on the neuron.
3. The only significant delay within the neural system is synaptic delay.
4. The activity of any inhibitory synapse absolutely prevents the excitation of the neuron at that time.
5. The structure of the network does not change with time.

Although these assumptions are some of the earliest in ANN development, they do not necessarily apply to current ANNs, especially the fifth assumption which is commonly replaced with the Hebbian Learning Law [105], allowing for changes in the structure of the network.

4.2.2. Simple perceptron

Arguably one of the first ANNs, simple perceptrons were developed by Frank Rosenblatt to act as electronic analogs to the human retina. Rosenblatt connected the input layer of the perceptron to a rectangular array of light sensors. A directed graph of a simple perceptron network is shown in Fig. 9(a). The weight connecting input node \( i \) to output node \( j \) is denoted by the matrix component \( w_{ij} \). The goal of Rosenblatt’s system was to trigger the relevant response unit given a class of input patterns. The input layer used a unipolar linear transfer function and the output was passed to the response layer by way of trainable weights.

The simple perceptron is said to use a form a supervised learning due to the fact that the weights were adjusted when an undesired response was triggered. The perceptron learning rule begins with a definition of the change made to the weight matrix component \( w_{ij} \) during the \( n \)th training set,

\[
\Delta w_{ij}^{(n)} = \alpha (d_j^{(n)} - z_i^{(n)}) x_i^{(n)} = \alpha e_j^{(n)} x_i^{(n)}, \tag{66}
\]

where \( d_j^{(n)} \) is the desired output value for input \( x_i^{(n)} \), \( z_i^{(n)} \) is the actual output from the \( j \)th neuron with input \( x_i^{(n)} \), \( e_j^{(n)} \) is the output error of the \( n \)th set and \( \alpha \) is referred to as the learning rate coefficient. Updating the weight matrix \( W \) proceeds as

\[
W^{(n+1)} = W^{(n)} + \Delta W. \tag{67}
\]

The result after \( n \) training sets is that each of the weights \( w_{ij} \) will have been updated according to the rule outlined in Table 5. The perceptron learning rule can be optimized through the tuning of the learning rate coefficient \( \alpha \). Small \( \alpha \) corresponds to a stable and slow learning scheme. Therefore, in practice, one would like to make \( \alpha \) as large as possible to ensure quick learning without introducing unstable oscillations about the desired value which result from over-relaxation.

In addition to the perceptron learning rule, Rosenblatt also proved a perceptron convergence theorem which states: given a finite set of inputs \( \mathbf{x} \) and desired output training sets \( \mathbf{d} \), each presented with a positive probability, the perceptron learning rule
provides guaranteed convergence of the weight matrix to values which give the correct outputs if and only if the said set of weights exists. For a detailed outline of the proof see Ref. [117]. However, there are many tasks for which a set of weights do not exists and, as mentioned before, in 1969 Minsky and Pappert published a book concerning the strengths and limitations of single layer perceptron networks [101]. Included in this book, and the less popular earlier text by E. B. Carne [118], was a criticism of the single layer perceptron’s inability to solve the simple two-state XOR problem which is described in Table 6. Solution of the XOR problem

<table>
<thead>
<tr>
<th>Classification</th>
<th>Expression</th>
<th>Derivative</th>
<th>$f(z)$</th>
<th>$f'(z)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unipolar step or heaviside</td>
<td>$f(y) = H(y) = \begin{cases} 1, &amp; \text{if } y &gt; 0 \ 0, &amp; \text{if } y &lt; 0 \end{cases}$</td>
<td>$\delta(y) = \begin{cases} 0, &amp; \text{if } y \neq 0 \ \infty, &amp; \text{if } y = 0 \end{cases}$</td>
<td><img src="image1" alt="Graph" /></td>
<td><img src="image2" alt="Graph" /></td>
</tr>
<tr>
<td>Bipolar step (threshold)</td>
<td>$f(y) = \text{sign}(y) - 2H(y) - 1$</td>
<td>$\delta(y) = \begin{cases} 0, &amp; \text{if } y \neq 0 \ \infty, &amp; \text{if } y = 0 \end{cases}$</td>
<td><img src="image3" alt="Graph" /></td>
<td><img src="image4" alt="Graph" /></td>
</tr>
<tr>
<td>Unipolar linear</td>
<td>$f(y) = \begin{cases} 0, &amp; \text{if } y &lt; -1 \ \frac{1}{2}(y + 1), &amp; \text{if }</td>
<td>y</td>
<td>&lt; 1 \ 1, &amp; \text{if } y &gt; 1 \end{cases}$</td>
<td>$\frac{1}{2}[H(y+1) - H(y-1)]$</td>
</tr>
<tr>
<td>Bipolar linear</td>
<td>$f(y) = \begin{cases} -1, &amp; \text{if } y &lt; -1 \ y, &amp; \text{if }</td>
<td>y</td>
<td>&lt; 1 \ 1, &amp; \text{if } y &gt; 1 \end{cases}$</td>
<td>$H(y+1) - H(y-1)$</td>
</tr>
<tr>
<td>Unipolar sigmoid (logistic)</td>
<td>$f(y) = \frac{1}{1 + e^{-y}}$</td>
<td>$f(y)(1 - f(y))$</td>
<td><img src="image9" alt="Graph" /></td>
<td><img src="image10" alt="Graph" /></td>
</tr>
<tr>
<td>Bipolar sigmoid (hyperbolic tangent)</td>
<td>$f(y) = \tanh(y)$</td>
<td>$(1 -</td>
<td>f(y)^2</td>
<td>)$</td>
</tr>
<tr>
<td>Gaussian radial basis</td>
<td>$f(y) = \exp(-</td>
<td>y-m</td>
<td>^2/\sigma^2)$</td>
<td>$-2(y-m)(f(y)/\sigma^2)$</td>
</tr>
</tbody>
</table>
would require a non-linear partitioning of even parity points from odd parity points. This is impossible for single layer perceptrons because they can only partition regions into two linearly separable spaces. As a result, researchers lost interest in neural computing and work on perceptrons was effectively wiped out for almost two decades.

4.2.3. Adaptive Linear Neuron (ADALINE)

Shortly after the conception of Frank Rosenblatt’s perceptron, Bernard Widrow and Ted Hoff developed the ADAPTive Linear NEuron (ADALINE) while at Stanford University. Like the TLU, the ADALINE is not a network, but rather a single logic neuron. Nonetheless, these artificial neurons can be connected in a single layer in much the same way as perceptrons (Fig. 9(a)). Unlike the unipolar step transfer function of the perceptron, ADALINES employed a bipolar step function (Fig. 3), which gave an output of ±1. Inputs of ADALINEs are also typically bipolar. The main contribution of Widrow and Hoff was the LMS training rule in which the weights linked with each input node are adjusted with an adaptive learning algorithm as follows: given a finite number \( N \) of training sets \( x_1, x_2, \ldots, x_N \) and desired outputs \( d_1, d_2, \ldots, d_N \), Widrow and Hoff defined a training cost given by

\[
J(\mathbf{w}) = E\left[ e_n^2 \right] \equiv \frac{1}{N} \sum_{n=1}^{N} e_n^2.
\]

where \( E[\cdot] \) indicates the expectation and \( e_n \) denotes the training error at the \( n \)th set, which is simply \( d_n - z_n \). Combining the definitions \( E[\mathbf{d}x] = \rho \) and \( E[\mathbf{x}x] = P \), where \( \langle \cdot, \cdot \rangle \) denotes the inner product, with the notation from above yields

\[
E\left[ e_n^2 \right] = E\left[ d_n^2 \right] = \mathbf{w}^T P \mathbf{w} - 2 \mathbf{w}^T \rho.
\]

Combining this result with (68), gives a gradient \( \nabla J \) which is used to minimize the cost function as

\[
\nabla J = \frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} = 2 \mathbf{P} \mathbf{w} - 2 \mathbf{\rho} = 0.
\]

Thus, the optimal Least Mean Square (LMS) setting of \( \mathbf{w} \) is given by

\[
\mathbf{w}_{\text{LMS}} = \frac{1}{P} \rho.
\]

Because the LMS rule for training makes use of expectations, the sample averages tend to be erroneous for a small number of \( N \) training sets. True convergence requires \( N \rightarrow \infty \). The simple ADALINE processing element has found applications in many engineering problems such as adaptive filtering, echo suppression, pattern recognition and prediction [117].

4.3. Multi-layer networks

4.3.1. MADALINE

In the 1980s, almost twenty years after Minsky and Papert’s criticism of the simple perceptron’s inability to solve the XOR problem, Bernard Widrow and Rodney Winter considered a new multi-layer network which consisted of Many ADALINEs (MADALINE) [119]. In addition to the input and output layers of simple ANN structures in Fig. 9(a), the MADALINE architecture included...
additional layers as in Fig. 9(b). These additional neural layers are termed hidden due to the fact that no partial outputs are available for the training of these nodes. The use of multiple layers is one way of circumventing the problem of computing mappings which require non-linear separability. It is the introduction of hidden layers that enabled the MADALINE architecture to solve the XOR problem and renew interest in neural computing. However, the history of the application of multi-layer ANNs to mappings which require non-linear separability actually began in 1900 with Hilbert’s formulation of his famous 23 problems for the coming century [120]. In particular, Hilbert’s 13th problem came in the form of the following supposition: there exist analytic functions of three variables that cannot be represented by a finite superposition of continuous functions of only two variables. This supposition was discredited by Kolmogorov in 1957 [121] whose reasoning was later refined by Sprecher in 1965 [122] who provided the following theorem [82].

For each integer \( n \geq 2 \), there exists a real, monotonically increasing function \( \psi(x) \) which is dependent on \( n \) and possesses the property: for each pre-allocated number \( \delta > 0 \) there exists a rational number \( \varepsilon \), \( 0 < \varepsilon < \delta \), such that every real continuous function of \( n \) variables \( f(x) \), can be represented by

\[
f(x) = \sum_{j=1}^{2n+1} \chi \left( \sum_{i=1}^{n} \lambda^i \psi \left( x_i + \varepsilon (j-1) \right) + j - 1 \right),
\]

where the function \( \chi \) is real and continuous and \( \lambda \) is a constant which is independent of \( f \).

The coupling of the previous theorem and multi-layer ANNs was pointed out by Hecht-Nielsen in 1987 [122] when he demonstrated that, as a result of the Sprecher theorem, any continuous mapping \( f \) can be regarded as a form of multi-layer ANN with two hidden layers. The output function of the first and second layers are given by \( \Psi \) and \( \chi \) respectively [122,123]. In addition to the work of Sprecher, Hornik et al. [124], Funahashi [123] and Cybenko [125] independently demonstrated that a general multi-layer ANN with sufficient hidden layers is adequate for the approximation of a continuous non-linear function on a finite closed interval.

 Widrow and Winter also developed a training algorithm known as the MADALINE adaptation Rule (MR). Widrow reasoned that the weights should be modified at each training set to reduce the error with as little disturbance to the representations learned by the previous training sets which he called the principle of minimal disturbance [117,119]. The MR training algorithm has been refined several times yielding procedures labeled MR, MII, MIII etc. The procedure outlined below is MIII.

The decision of whether or not the weights for a given layer should be adjusted requires an input vector \( \mathbf{x} \) and the appropriate target output \( \mathbf{d} \) in order to calculate the sum of the squared output errors. Widrow changed the input to the \( k \)th neuron by some small amount \( \Delta s \) and investigated the change in the sum squared output error given by

\[
\Delta e^2 = \Delta \left( \sum_{k} e_k^2 \right).
\]

The gradient of the sum squared output error term with respect to its weight vector is estimated through the use of finite differences as

\[
\nabla_k = \frac{\partial (e_k^2)}{\partial s_k} = \frac{\Delta (e_k^2)}{\Delta s} x_k.
\]

which is subsequently used to reduce the error in a direction directly opposite of the gradient;

\[
w_{k+1} = w_k - \alpha \frac{\Delta (e_k^2)}{\Delta s} x_k.
\]

Finally, defining \( f = f(1-f) \) as the derivative of the sigmoid function (see Table 3), the MIII training rule given is by

\[
w_{k+1} = w_k + 2\alpha e_k f'(s_k) x_k.
\]

where \( \alpha \) is again the learning rate coefficient. It should be noted that this training algorithm is similar to the back-propagation training algorithm discussed in the next section. MADALINES have found applications in many areas outside of solar irradiance modeling and forecasting, such as invariant pattern recognition, missile guidance and detonation and general computations of well-behaved functions [117].

4.3.2. Multi-layer perceptron

Multi-Layer Perceptrons (MLPs) with Back Propagation-Learning (BPL) are some of the most established ANN architectures due to their ability to perform arbitrary non-linear mappings. MLPs are also sometimes referred to as Multi-Layer Feed-Forward (MLFF) networks to emphasize the forward flow of information with respect to the backward direction of weight adjustment used in BPL algorithm. The structure of MLPs are similar to MADALINES, with layered neurons possessing only forward connections to successive layers (see Fig. 9(b)). However, MLPs differ from MADALINES in the derivation of the popular BPL algorithm outlined next. As we have seen before, the LMS and Perceptron Learning rules perform in very similar ways. Nonetheless, neither of these learning rules can be applied to MLPs because these methods do not identify how to make adjustments to weights associated with hidden-layers. As a result, a new method using upstream variables was developed.
Several authors independently derived the BPL method in the 1970’s and 1980s. Paul J. Werbos proposed the BPL method first in his 1974 Harvard University doctoral dissertation [126]. Almost ten years later, in 1985, D. B. Parker re-derived the BPL method in his MIT technical report [127]. Even so, credit for developing the BPL algorithm into a realizable procedure is typically assigned to David Rumelhart and the other members of his distributed processing group at the University of California, San Diego, in 1985 [128].

In order to begin a discussion of BPL, we consider an MLP with a single hidden layer which has been modified for the BPL algorithm as shown in Fig. 10. The results can then be generalized to an MLP with an arbitrary number of hidden layers. As before, we adopt the following convention for weight indexing: the synaptic weight connecting input neuron $i$ to hidden layer neuron $j$ is denoted by $w_{ij}$ and the weight connecting hidden layer neuron $j$ to output neuron $k$ is denoted by $w_{kj}$. We will also define the net input to the hidden and output layers’ neurons as,

$$H_j = \sum_i w_{ij} x_i$$

and

$$O_k = \sum_j w_{kj} z_j$$

respectively. This is similar to an unbiased version of Equation (64). Using the above notation we can write the outputs of node $j$ of the hidden layer and unit $k$ of the output layer as

$$z_j = f(H_j)$$

$$z_k = f(O_k)$$

In order to begin training one must have differentiable transfer functions, such as the sigmoid function, and access to a training data set consisting of $N$ pairs of input patterns. As usual, a training method is developed in order to reduce the total system error for all training patterns through an adjustment of the weights. The total system error $E_{sys}$, defined as the average of the output errors over all training patterns, is

$$E_{sys} = \frac{1}{N} \sum_{n=1}^{N} E_n.$$

As has been done before, weights are adjusted on each successive pattern presentation proportional to the negative of the error gradient. Therefore, at the $k + 1$ step we can write

$$\Delta W^{(k+1)} = -\alpha \frac{\partial E_n}{\partial W^{(k)}},$$

where $\alpha$ is the familiar learning rate coefficient. The total system error can be written as

$$\frac{\partial E_{sys}}{\partial W} = \frac{1}{N} \sum_{n=1}^{N} \frac{\partial E_n}{\partial W}$$

Here we define $E_n$ as the mean square error,

$$E_n = \frac{1}{2} \sum_{n=1}^{N} (d_n - z_n)^2.$$

In order to express the update rule in terms of the system parameters, we must successively apply the chain rule. Since we know for output unit $k$

$$\frac{\partial E}{\partial z_k} = -(d_k - z_k)$$

and

$$\frac{\partial z_k}{\partial l_k} = f'(l_k),$$

we can write

$$\frac{\partial E}{\partial l_k} = \frac{\partial E}{\partial z_k} \frac{\partial z_k}{\partial l_k} = -(d_k - z_k)f'(l_k).$$

By defining the previous operation as a generalized delta term,

$$\delta_k = (d_k - z_k)f'(l_k),$$

we can now write an update rule for the output units as

$$\Delta w_{jk} = -\alpha \frac{\partial E}{\partial w_{jk}} = \alpha \delta_k y_j.$$  

The same reasoning is applied to the hidden layer weights. We want an expression for the update rule given by

$$\Delta w_{ij} = -\alpha \frac{\partial E}{\partial w_{ij}} = -\alpha \frac{\partial E}{\partial H_j} \frac{\partial H_j}{\partial w_{ij}}.$$  

Examination of the first term reveals that another application of the chain rule is required,

$$\frac{\partial E}{\partial H_j} = \frac{\partial E}{\partial z_j} \frac{\partial z_j}{\partial H_j} = \frac{\partial E}{\partial z_j} f'(H_j),$$

while the second term can be written explicitly from Equation (77) as,

$$\frac{\partial H_j}{\partial w_{ij}} = \sum_i \frac{\partial}{\partial w_{ij}} (w_{ij} x_i) = x_i.$$  

Using the fact that we can express the output from node $z_k$ for a given input pattern as

$$z_k = f(l_k) = f \left( \sum_j w_{kj} z_j \right) = f \left( \sum_j w_{kj} f(H_j) \right)$$

along with the error defined in Equation (84), we are able to directly differentiate the $\partial E/\partial z_j$ term in Equation (91), obtaining

\begin{align*}
\text{\textbf{Fig. 10.}} \text{ Modification of an MLP network with a single hidden layer for the application of the BPL algorithm. The use of partial derivatives of errors with respect to the weights upstream of the flow of information allows the algorithm to overcome the problem of inaccessible of hidden layer data and is at the core of the BPL. The resulting algorithm may then be generalized to an MLP with an arbitrary number of hidden layers.}
\end{align*}
\[
\frac{\partial E}{\partial z_j} = \frac{1}{2} \sum_k \left( d_k - f\left(\sum w_{jk} z_k\right)\right)^2 \tag{94}
\]
\[
= -\sum_k (d_k - z_k) f'(h_k) w_{jk}. \tag{95}
\]

As done before, the previous expression may be conveniently written as a general delta operator
\[
\delta_j = f'(h_j) \sum_k \delta_k w_{jk}, \tag{96}
\]
and the update rule for the hidden layer can be written as
\[
\Delta w_{ij} = \alpha \delta_j x_i, \tag{97}
\]
which is very similar to the perceptron learning rule summarized in Table 7. Turning our interest to the hidden layer nodes, we do not have target values to use in the computation of an error like we did with the output layers. As a result, in this case, we must distribute the output errors in some way in order to adjust the weights connecting input node \( i \) to hidden node \( j \). This is accomplished through the use of \( \delta_j \)
\[
\delta_j = f'(h_j) \sum_k \delta_k w_{jk}, \tag{98}
\]
which was discussed earlier. Thus, one can see that the BPL algorithm takes partial derivatives of errors with respect to the weights upstream, which allows it to overcome the inaccessibility of hidden layer data. The resulting errors are then propagated “backwards” through the network every time a training pattern is presented. For a detailed derivation of the BPL algorithm see Refs. [82,116,117,126–128].

4.5. k-nearest neighbors (k-NN)

Despite the similarity in the acronym, the k-nearest neighbors (k-NN) method is unrelated to ANNs. Although both methods can be considered Artificial Intelligence techniques, the origin of the k-NN stems from classification theory and its origins can be tracked back to an unpublished technical report by Fix & Hodges regarding the density estimation for classification theory. Even though their work has not been published in peer-review literature Fix & Hodges have received ample recognition in classification theory textbooks and articles [144–146].

As an algorithm for pattern classification in classification theory the k-NN method is one of several nonparametric techniques for estimating the density functions from sample patterns. Within the confines of classification theory the k-NN has a very specific place and its mathematical derivation and properties can be derived precisely from the underlying theory. Outside of the confines of classification theory the k-NN is often defined more heuristically by the following statement (or some version thereof): an initial set of objects is created and each object is classified by a property or value; new objects are classified based on the decision rule that assigns to an unclassified object the class/value of the nearest \( k \) objects (hence the designation k-nearest neighbors), where \( k \) is a positive integer, typically small. If \( k = 1 \), then the object is simply assigned to the class of its nearest neighbor. Fig. 11 depicts this heuristic definition graphically.

The k-NN is one of the simplest methods among the machine learning algorithms. In contrast to the statistical methods presented before that attempt to find models from the available data, the k-NN uses the training data as the model, and unlike the ANN models there are no unknown model coefficients that need to be determined through training. Despite the fact that the k-NN model was originally developed for pattern classification its simplicity allows it to be applied to regression problems for time-series [147], such as the forecast of solar radiation. Although there are very few articles in literature that apply k-NN to the forecasting of solar irradiation (see for instance [138,148]), k-NN has been extensively applied as a forecasting technique to problems such as: the forecast of electricity load and electricity market price [149,150], the
forecast of daily river water temperature [151], the forecast of water inflow [152] and weather forecast [153], just to cite some of the most recent publications. In the field of meteorology and weather forecasting the k-NN method is also known as the analog method [154–156]. The analog method searches a database of past weather patterns that resemble most closely the current weather conditions. The patterns are described either in terms of circulation patterns or in terms of various surface weather parameters. The selected past weather conditions, also known as analogs, are then used for predicting weather and its various components such as maximum, minimum and ambient temperatures, average wind speed, atmospheric pressure, etc. Likewise, for the propose of forecasting time-series, the k-NN model consists of looking into the time-series history and identifying the timestamps in the past that resemble the “current” conditions most closely – the nearest neighbors. Once they are determined the forecasting is computed from the time-series values subsequent to the matches. In essence, the k-NN model resembles a lookup table for which previous patterns can be obtained as:

\[ z_{t + \Delta t} = \sum_{i=1}^{k} a_i z_{t_i + \Delta t} / \sum_{i=1}^{k} a_i \]  \hspace{1cm} (101)

In other words, the prediction \( z_{t + \Delta t} \) is computed from a linear combination of the time-series values that follow the nearest neighbors \( t_i \). The weights \( a_i \) often are a function of the distance \( d_j - d_j \) – Fig. 11.

5. Remote sensing models

Solar forecasts are also used in engineering applications such as the design and operation of solar power plants and the modeling of agricultural variables including vegetation and crop growth, crop yield, soil moisture, surface energy and moisture balance [16,50,157–160]. However, all of the solar forecasting methods discussed so far require accurate measurement of the solar resource. Dissatisfyingly, precise pyranometer measurements have historically been rangebound to scattered weather stations and universities due to the expense and labor required to procure, maintain and quality check instruments and data. For this reason, when no data was available for a given location it was common to extrapolate data from the nearest ground measurement station [161]. This approach is appropriate for sites in proximity to measurement stations to constitute strong micro-climatic and environmental correlation. Having said that, it is a well known result.
that the error of this method increases with extrapolation distance and temporal refinement [161] thereby limiting its application. An alternative to a national or global scale ground sourced monitoring network is provided through satellite based irradiance measurements of the Earth and its atmosphere.

Origins of the meteorological satellite can be traced back to a paper published in 1919 by Robert H. Goddard titled “A Method of Reaching Extreme Altitudes” [162]. A decade later, in 1929, Goddard launched a liquid fueled rocket carrying instrumentation which included a thermometer, a barometer and a camera [163]. Advancements in rocket technology over the two decades following Goddard’s work led to the first images of clouds from high altitudes [164,165] and eventually the launch of Sputnik-1; the first artificial satellite to be put into Earth’s orbit by the Soviet Space Program. Shortly after, in 1958, the U.S. formed the National Aeronautics and Space Administration (NASA) which would launch a number of meteorological satellite missions; many of which would carry instrumentation useful for atmospheric modeling.

5.1. Orbits

Before discussing the assortment of satellite based methods it is useful to address some foundational principles. There are two basic types of orbits which a meteorological satellite may occupy: Low Earth Orbit (LEO) and GeoStationary Orbit (GSO). Each type of orbit presents its own set of advantages and disadvantages. LEOs get their name from their relatively low altitude of approximately 750–850 km or about 10% of the radius of the Earth. This low orbiting altitude implies a short period, typically on the order of 90–100 min, which leads to about 15 orbits per day. Satellites in a LEO can be used to measure ozone concentrations or atmospheric temperatures among many other quantities.

LEOs may be further classified as sun-synchronous, polar or both. A satellite in a Sun-Synchronous Low Earth Orbit (SSLEO) integrates altitude and inclination in such a way that the satellite passes over a given location on the earth at the same local solar time. For example, a satellite in an SSLEO might cross over the equator fifteen times a day each time at 11:00 a.m. local time. Satellites in an SSLEO also need to process approximately one degree per day to keep up with the revolution of the Earth around the sun. SSLEOs are used for satellites which require constant illumination for imagery. Satellites that make use of visible band radiation are placed in SSLEOs with bright sunlight while satellites which measure long-wave radiation are placed in darkness.

A satellite in a Polar Low Earth Orbit (PLEO) has an inclination of almost 90° and passes over (or near) the planet’s poles on each revolution. PLEOs remain fixed with respect to the rotation of the Earth which will cause the satellite to pass over the equator at a different longitude on successive orbits. This allows the satellite to image many different parts of the Earth’s surface as it rotates underneath the satellite. The shortcoming of these orbits is that they image different parts of the Earth at different times making observation of one location difficult. The exception to this is if the satellite is in what is known as a Polar Sun-Synchronous Low Earth Orbit (PSSLEO). Satellites in a PSSLEO may be used to conduct high resolution imagery and radiometry of a single location a finite number of times per day.

GSOs are orbits in which the satellite circles the Earth at the same rate as the Earth spins which places them at an altitude of approximately 36,000 km or about one tenth of the distance to the moon. Satellites in GSO are located near the equator so there is a constant force of gravity due to the bulging of the Earth at this latitude. Consequently, the satellite appears stationary over a certain point on the equatorial plane which allows frequent full disk observations. However, because the satellites are almost fifty times farther than their LEO counterparts the image resolution is typically much lower, on the order of 3–5 km. Fig. 13 illustrates some of the fundamental differences between LEOs and GSOs. Once the various types of meteorological satellite orbits are understood we can move on to a discussion of the specific satellite missions.

5.2. Satellites

NASA launched its first LEO meteorological satellite, the Television InfraRed Observation Satellite (TIROS-1), on the first of April, 1960. TIROS-1 was equipped with a five-channel radiometer designed by Hanel [166]. The radiometer’s third channel measured solar radiation reflected by Earth and its atmosphere in a spectral range of 0.2–6 μm. TIROS-1 would be succeeded by nine other similar satellites (TIROS 2–10) launched between 1960 and 1966.

Shortly after the initial TIROS program, NASA developed the Environmental Science Services Administration (ESSA) satellite program which lasted four years and included the launch of nine satellites. ESSA was essentially an extension of the mission objectives from TIROS; namely to provide cloud-cover photography to national meteorological centers. Resolution was improved over the TIROS program and ESSA satellites included a “cartwheel” feature initially tested on TIROS-9. This cartwheel configuration combined with two cameras mounted 180° opposite each other and an SSLEO allowed the ESSA series to image a given point on the surface every time the satellite rotated about its axis. However, radiometers were deployed only on ESSA-7 and ESSA-9.

As a result of the success of the early ESSA missions the NIMBUS program was initiated in 1964 and included seven satellites. The NIMBUS satellites were launched into SSLEO and used a new automatic picture transmission to more than 60 low-cost data acquisition stations on the ground. The Advanced Vidicon Camera System (AVCS), which was used for daylight imaging, and the infrared Flat Plate Radiometer (FPR), which was employed to monitor night time cloud cover conditions, were contributed to the ESSA-9 mission. In addition, the NIMBUS program allowed for the first global radiation budget which permitted the refinement and validation of climatological models. Other LEO satellites include NASA’s Aqua and Terra series used to monitor the Earth’s water cycle and changing climate respectively [167].

Current operational LEO satellites include the U.S. National Oceanic and Atmospheric Administration (NOAA) series, also known

![Fig. 13. Illustration of a Polar Sun-Synchronous Low Earth Orbit (PSSLEO) satellite and a Geostationary Orbit (GSO) satellite. The PSSLEO satellite has an altitude which is nearly 50 times lower than the GSO and an inclination of nearly 90° with respect to the GSO and equatorial plane. As a result, the PSSLEO satellite completes approximately fifteen orbits per day tracing out the disk shown in the image while the GSO satellite remains fixed above a single point on the equator. The PSSLEO satellite is able to conduct high resolution imagery and radiometry of a location of interest only a finite number of times per day. On the other hand, the GSO satellite can provide continuous full disk observations, but at a substantially reduced resolution.](image-url)
as the Advanced TIROS-N (ATN), and the European Organization for the Exploration of Meteorological Satellites (EUMETSAT) MetOp series [167]. The NOAA series was initially developed as the Improved TIROS Operational System (ITOS) and then the Next generation TIROS (TIROS-N) programs which were launched between 1970 and 1994. Both the NOAA and MetOp series of LEO satellites are equipped with the third generation of the visible/infrared Advanced Very High Resolution Radiometer (AVHRR) and the Advanced Microwave Sounding Unit (AMSU) [167]. Other LEO satellite programs include Russia’s Meteor satellite series and China’s Feng Yun satellites. In October 2011 NASA’s Suomi National Polar-observing Partnership (NPP) launched a new satellite which carried five instruments: the Advanced Technology Microwave Sounder (ATMS), the Cross-track Infrared Sounder (CrIS), the Ozone Mapping and Profiler Suite (OMPS), the Visible Infrared Imaging Radiometer Suite (VIIRS) and Clouds and the Earth’s Radiant Energy System (CERES).

While LEO satellites are well suited for the high resolution monitoring of specific locations at finite intervals, continuous full disk observations were required for applications of weather analysis. Accordingly, interest shifted from PLEO to GSO satellites. In 1966 and 1967, two experimental Applications Technology Satellites (ATS–1 and ATS–3) were launched by NASA in order to examine the practicality of placing a satellite into a GSO. The ATS series was primarily intended for communications, but also carried meteorological related equipment including a spin-scan cloud camera. Shortly after, in 1974 and 1975, two more GSO satellites were launched by NASA and operated by NOAA, Synchronous Meteorological Satellites (SMS-1 and SMS-2). The SMS series carried a visible infrared spin-scan radiometer (VISSR) for day/night measurements of cloud cover and temperatures of the earth and atmosphere.

The high-altitude GSO techniques developed during the ATS and SMS programs paved the way for current operational GSO satellites which include the NOAA Geostationary Operational Environmental Satellite (GOES), the European METEOSAT series and the Japanese Multifunctional Transport Satellites (MTS) [167]. The instrumentation on these operational GSO satellites tend to vary by operator, however the capabilities of the instrumentation share a number of characteristics including the ability to measure visible and infrared radiation. Typically, images are acquired every 30 min although the METEOSAT’s Spinning Enhanced Visible and Infrared Imager (SEVIRI) possesses a nominal sampling frequency of 15 min [167,168] and the GOES imager can scan small areas of interest in as little as 30 s [167,169].

5.3. Radiation budget

Prior to the launch of TIROS–1 it was impossible to make direct measurements of the radiation that originated from the sun and was subsequently scattered back into space by the earth and its atmosphere [170]. This scattered fraction is referred to as the planetary albedo (not to be confused with the atmospheric or terrestrial albedo discussed later). Previous attempts to quantify planetary albedo made use of indirect measurements of the dark side of the moon, which was consequently illuminated by a fraction of the sunlit earth [171–173]. Other approaches involved the estimation of cloud distribution over the planet and the assignment of specific reflectivities to various cloud types [174–176].

Eventually, satellites equipped with radiometers allowed for direct measurement of the planetary albedo. Fig. 14 illustrates the fundamental relationship between satellite observation of planetary albedo and ground level irradiance. It should be noted that a slightly different notation is used in this section in order to stay consistent with notations commonly used in the field. By the conservation principle, the difference between the net irradiance at the top of the atmosphere $I_{TOA}^\text{in}$ and the net irradiance at the ground level $I_{in}^\text{out}$ must be equal to the flux lost (either reflected, scattered or absorbed) in the volume joining the two.

$$I_{TOA}^\text{in} - I_{in}^\text{out} = I_{in}^\text{TOT} + I_{out}^\text{TOT}.$$ (102)

In the previous equation $I_{in}^\text{TOT}$ is the flux lost due to absorption in the atmosphere and $I_{out}^\text{TOT}$ is the total irradiance reflected (or scattered) back out of the atmosphere into space, a combination of the irradiance reflected from the ground $I_{out}^\text{ground}$ and the irradiance reflected from the atmosphere $I_{out}^\text{atmosphere}$:

$$I_{out}^\text{TOT} = I_{out}^\text{ground} + I_{out}^\text{atmosphere}.$$ (103)

Substitution of (103) into (102) combined with normalization with respect to $I_{TOA}^\text{in}$ and a separation of atmospheric and terrestrial terms yields,

$$\frac{I_{TOA}^\text{in}}{I_{TOA}^\text{in}} + \frac{I_{out}^\text{out}}{I_{TOA}^\text{in}} = 1 - \frac{I_{out}^\text{ground}}{I_{TOA}^\text{in}} \frac{I_{out}^\text{atmosphere}}{I_{TOA}^\text{in}}.$$ (104)

The right hand side of the previous equation can be written as,

$$\frac{I_{TOA}^\text{in}}{I_{TOA}^\text{in}} + \frac{I_{out}^\text{out}}{I_{TOA}^\text{in}} = \left(1 + \frac{I_{out}^\text{ground}}{I_{out}^\text{atmosphere}}\right) \frac{I_{in}^\text{ground}}{I_{TOA}^\text{in}},$$ (105)

which gives the following relationship

$$\frac{I_{TOA}^\text{in}}{I_{TOA}^\text{in}} = \frac{1 - \frac{I_{out}^\text{ground}}{I_{out}^\text{atmosphere}} \frac{I_{out}^\text{atmosphere}}{I_{TOA}^\text{in}}}{1 + \frac{I_{out}^\text{ground}}{I_{out}^\text{atmosphere}} \frac{I_{out}^\text{atmosphere}}{I_{TOA}^\text{in}}}.$$ (106)

The observant reader may recognize the clearness index $K_t$ on the left hand side of the equation along with the atmospheric absorptivity,

$$\alpha_a = \frac{I_{out}^\text{ground}}{I_{TOA}^\text{in}};$$ (107)

the atmospheric albedo,

$$\rho_a = \frac{I_{out}^\text{atmosphere}}{I_{TOA}^\text{in}};$$ (108)
and the terrestrial albedo,
\[ \rho_g = \frac{g_{\text{out}}}{g_{\text{in}}}. \]  \hspace{1cm} (109)

Employing the previous definitions reveals the relationship between clearness index and quantities which can be measured by the satellite’s radiometers in a compact form,
\[ K_t = \frac{1 - \alpha_d - \rho_g}{1 + \rho_g}. \]  \hspace{1cm} (110)

As one would expect, the atmospheric albedo intensifies with increasing atmospheric turbidity and cloudiness. In addition, heightened atmospheric albedo also implies the attenuation of the radiation reaching the ground. If we are able to estimate the atmospheric absorption coefficient and know the terrestrial albedo, a priori, we would be able to approximate the clearness index \( K_t \) from the measured satellite data using Equation (110).

The method used to approximate the aforementioned parameters typically varies by author. Nonetheless, the relationship between the clearness index and the atmospheric and terrestrial parameters in Equation (110) is at the foundation of all satellite based irradiance forecasting models. Measurement of \( g_{\text{in}} \) in conjunction with the knowledge of the Linke turbidity coefficient, for example, and a priori knowledge of \( g_{\text{out}} \) (from historical measurements of clear day planetary albedo) allows for the estimation of ground level irradiance from information regarding the clearness index \( K_t \).

Depending on the treatment on the interaction of solar irradiance and the atmosphere satellite models may be classified into two categories: physical models and statistical models [177,178]. The benefits of physical models include their generality and the elimination of the requirement of ground data through the use of RTMs. Correspondingly, RTMs introduce the requirement of precise and comprehensive measurement of the atmospheric morphology as well as careful calibration of the satellite data. Statistical models, on the other hand, rely on straightforward statistical regressions between satellite and ground measurements. As a result, statistical models are less computationally expensive than RTM based models, however, they suffer from their loss of generality and requirement of ground data.

5.4. Physical satellite models

Physical satellite models for solar irradiance forecasting are based on the interactions between solar radiation and participating atmospheric components such as gasses (\( \text{CO}_2, \text{H}_2\text{O}, \text{N}_2, \text{O}_2, \text{O}_3, \text{etc.} \)) and aerosols. Similar to a number of the clear sky models covered in Section 2.3, the physical interactions are typically modeled by way of RTMs. Thus, physical satellite models can be thought of as an improvement upon RTM based clear sky models through the addition of information regarding current atmospheric conditions. Atmospheric conditions are accounted for through the measurement of local meteorological data. This eliminates the need for solar irradiance data at the surface, but, because these models need to convert digital counts from satellite based radiometers into a corresponding flux densities, accurate and frequent calibration of the instrument is required [178].

5.4.1. Gautier-Diak-Masse model

One of the earliest physical models was developed by Gautier, Diak and Masse (GDM) in 1980 [179]. The GDM model considered clear and cloudy conditions separately. The determination of whether a given pixel was clear or cloudy was accomplished by selecting a brightness threshold obtained through the selection of the minimum value at every pixel for every hour from the past several days. One shortcoming of the original GDM model was the absence of variations in terrestrial albedo with changing solar zenith angle. Raphael and Hay [180] included the T-minimum brightness determination [177] in order to correct for the previous consideration.

The GDM clear sky RTM included several input parameters: reflection coefficient for diffuse radiation, atmospheric transmissivity as a function of the reflection coefficient for DNI, the absorption coefficient for slant water vapor path and the solar zenith angle, the atmospheric albedo as a function of the irradiance received by the satellite, and a second absorption coefficient for slant water vapor path for the satellite zenith angle. Reflection coefficients were calculated using the results from Coulson [181,182] while the absorption coefficients made use of the expressions from Paltridge [183]. Improvements to GDM model include the absorption of ozone and aerosols by Gautier and Frouin [184] in addition to multiple reflections by Tanrè et al. [185].

The cloudy sky RTM allowed for the simple treatment of stratiform low and middle altitude clouds [179]. Absorption above and below the clouds were considered separately and further classified as either upwelling or downwelling. Similar inputs for the cloudy sky RTM were considered: cloud albedo as a function of the absorption of short wave radiation above and below the cloud, and cloud absorption coefficient estimated on the basis of the satellite’s measurement of the visible brightness of the cloud. The authors assumed a simple linear relationship between measured visible brightness and absorption ranging from zero for no cloud to a maximum of 0.2 for very bright clouds [179].

5.4.2. Marullo-Dalu-Viola model

Marullo, Dalu and Viola (MDV) re-evaluated the GDM model in 1987 using data for the METEOSAT data for the Italian peninsula [186]. Like its predecessor, the MDV model also considered clear sky and cloudy conditions separately. However, the terminology “standard atmosphere” and “real atmosphere” were used in place of clear sky and cloudy sky respectively.

The MDV “standard atmosphere” model was similar to the GDM clear sky model except for the addition of a reflecting non-absorbing layer which accounted for the presence of aerosols in the atmospheric column. The MDV RTM was adapted from Schmetz [187] and required information regarding the temperature profile of the atmosphere, water vapor content and a three-layer aerosol column which describes regional clear sky conditions. However, like the GDM model the MDV model did not account for the variation of surface albedo with solar zenith angle. Planetary albedo for a standard atmosphere was approximated through the use of regional clear sky data and assumed to be uniform for the region, varying only with solar zenith angle.

Any significant deviation from the standard atmosphere model was assumed to be a consequence of atmospheric particle loading. The atmospheric loading in the real atmosphere was resolved by a thin reflecting non-absorbing layer assumed to be higher than the particles responsible for scattering in the standard atmosphere.

5.4.3. Möser-Raschke model

Möser and Raschke (MR) also used METEOSAT images to estimate ground level irradiance [188], but the authors used the RTM developed by Kerschgens [189] which was more complex than the previous models. Improvements to the MR method include the addition of more parameters to accurately describe the state of the atmosphere and infrared data which is used to estimate the cloud top height.

Input parameters include the solar zenith angle, cloud top height, optical depth of the clouds, terrestrial albedo, boundary
layer structure, climatological profiles of temperature, pressure, humidity, ozone concentration and cloud droplet size distribution. One significant result of this model was the demonstration that clouds, rather than aerosols, have the greatest impact on irradiance reaching ground level.

5.4.4. Dedieu-Deschamps-Kerr model

Unlike the previous modes, the model developed by Dedieu, Deschamps and Kerr (DDK) in 1987 [190] used a single equation valid for both clear and cloudy conditions. This was accomplished by considering a clear sky model in conjunction with a model which considered only the effect of clouds on solar irradiance. The assumption was made that the other effects, such as aerosol loading was constant.

Input parameters include a sky transmissivity factor, which accounted for Mie and Rayleigh scattering as well as gaseous absorption using the formulae of Lacis and Hansen [191] together with the RTM of Tanrè [185], and planetary and terrestrial albedo determined from with the METEOSAT radiometer data. Multiple reflections between the cloud base and the ground were assumed to behave isotropically. It should be noted that as a consequence of uniformity of the aerosol content in both the clear sky and cloudy conditions the model treats an unusually strong concentration of aerosols as a cloud [177].

5.5. Statistical satellite models

Statistical satellite models are based on regressions between solar irradiance measured by pyranometers at the ground level and simultaneous digital counts provided by satellite based instrumentation. Rather than use RTMs, a number of independent parameters are included in various regression equations: solar zenith angle, cloud cover index, atmospheric transmissivity, current brightness, minimum brightness and maximum brightness of each pixel. As pointed out in Ref. [177] there are two main difficulties which arise when comparing satellite and ground data. The first is given by the errors associated with the localization of the ground based pyrometer sites on the satellite images. The second arises from the fundamental difference in the measurement technique: satellite data are instantaneous “snapshots” over a small solid viewing angle while ground data have been historically integrated over time (typically an hour) over a large solid viewing angle (2π). Several authors have suggested solutions to these problems, including the incorporation of more pixels in the definition of target areas as well as enhancing satellite resolution.

5.5.1. Hay-Hanson model

One of the simplest statistical satellite models was developed by Hay and Hanson (HH) in 1978 [192]. The model was developed for the Global Atmospheric Research Program’s Atlantic Tropical Experiment to generate maps of the shortwave radiation (0.55–0.75 μm) reaching the surface of the ocean. The HH model is based of a statistical linear regression of the clearness index and atmospheric absorptivity:

\[ \text{Kt} = a - b_0 q_a. \]  

Substituting Equations (7) and (111) into (6) and solving for \( k_t \) yields,

\[ k_t = \text{Kt} d_0 \cos(\theta_l) = (a - b_0 q_a) d_0 \cos(\theta_l). \]

Hay and Hanson [192] originally determined regression coefficients \( a \) and \( b \) as

\[ a = 0.79 \quad b = 0.71 \]

These values were later re-evaluated by Raphael and Hay [180] to be

\[ a = 0.788 \quad b = 1.078 \]

which gives a better agreement with their dataset.

It has been pointed out [177] that this relationship fails under unusually high surface albedo which results from a snow- or ice-covered surface. In addition, despite what has been mentioned about statistical methods, this approach requires the calibration of reported digital satellite counts \( i \) in order to determine visible radiance.

5.5.2. Tarpley & Justus-Paris-Tarpley models

Tarpley used a set of coincident satellite and ground pyrometer data sets taken by the National Environmental Satellite Data and Information Services (NESDIS) and the Great Planes Agricultural Council over the central U.S. in late 1970s [193]. This study made use of statistical regressions against measurements from GOES VISSR. Three separate cases were considered based on the value of the cloud index defined by Tarpley as

\[ n = 0.5 N_2 + N_3 \frac{1}{N}, \]  

where \( N \) is the total number of pixels included in the target area, and \( N_2 \) and \( N_3 \) are the number of pixels in partly cloudy and cloudy conditions respectively. The Tarpley regression model was defined as

\[ I = \begin{cases} 
 a + b \cos(\theta) + c K_t + d n + e \left( \frac{B_{\text{cl}}}{B_{\text{m}}} \right)^2 & \text{if } n \leq 0.4 \\
 a + b \cos(\theta) + c \left( \frac{B_{\text{cl}}}{B_{\text{m}}} \right)^2 & \text{if } 0.4 < n < 1. \\
 a + b \cos(\theta) + c \left( \frac{B_{\text{cl}}}{B_{\text{m}}} \right) & \text{if } n = 1 
\end{cases} \]  

where \( B_m \) is the mean target brightens, defined as the mean brightness of a 7 × 6 pixel array, \( B_{\text{cl}} \) is the mean cloud brightness, estimated through an average of the brightness values of all the pixels in the target area brighter than a specified threshold, and \( B_{\text{cl}} = B_{\text{cl}}(\theta = 45^\circ, \phi_i = 105^\circ) \) is the normalized clear brightness which is a special case of the clear brightness \( B_0 \) obtained from the following regression,

\[ B_0 = a + b \cos(\theta) + c \sin(\theta) \cos(\phi_i) + d \sin(\theta) \cos^2(\phi_i). \]

Raphael and Hay [180] also estimated their own regression coefficients for this model which are different from Tarpley’s treatment (see Table 8).

This model was later refined by Justus, Paris and Tarpley (JPT) [194] for part of the Agriculture and Resources Inventory Surveys through Aerospace Remote Sensing (AgRISTARS) program. This new model replaced the three equations of Tarpley’s model with the following single equation,

\[ I = I_0 \left( \frac{R_{\text{cl}}}{R_{\text{m}}} \right)^2 \cos(\theta) \left[ a + b \cos(\theta) + c \cos^2(\theta) \right] + d \left( B_{\text{cl}}^2 - B_0^2 \right), \]

where \( B_{\text{cl}} \) is again the mean observed target brightness and \( B_0 \) is defined by the following relationship,
As before, the authors in Ref. [194] assumed that the brightness for clear sky conditions were known. The weights $w_1$ and $w_2$ are values between 0 and 1 which were empirically determined. Each of the cases above approximates various conditions of the atmosphere. The first and fifth cases correspond to the likely presence of clouds and the insufficient scene illumination for radiation forecasts respectively. Each of these cases leaves the clear brightness unaltered. The second case allows for seasonal variation in the clear brightness due to snow- or ice-cover. The third case is to account for clearer than normal days while the fourth case allows for the removal of erroneous scene brightness for each pixel [16].

The model relates hourly global irradiance $I_t$ and cloud index $n_t$ through a simple regression:

$$I_t = I_{clrt}(nt) [a + b]$$  \tag{118}

where $f(nt)$ is a fifth order polynomial of the cloud index given by

$$f(nt) = c_5n_t^5 + c_4n_t^4 + c_3n_t^3 + c_2n_t^2 + c_1n_t + c_0.$$  \tag{119}

Values of the coefficients, as calculated by Perez, are given in Table 9. This model was also modified by Perez and Ineichen to forecast DNI from GHI forecasts provided by the operational model as well as corrections for locations presenting complex arid terrain [202].

### 5.5.4. Perez Operational model

One of the most widely used statistical satellite models is the operational model of Perez [201]. The Perez model uses a modified version of Kasten’s clear sky model (see Section 2.3.8) which defines a Linke turbidity coefficient independent of airmass [10]. The model also allows for the modification of the algorithm based on real time measurements of snow- or ice-cover as well as the correction of sun-satellite angle effects for each pixel [16].

The model relates hourly global irradiance $I_t$ and cloud index $n_t$ through a simple regression:

$$I_t = I_{clrt}(nt) [a + b].$$  \tag{118}

where $f(nt)$ is a fifth order polynomial of the cloud index given by

$$f(nt) = c_5n_t^5 + c_4n_t^4 + c_3n_t^3 + c_2n_t^2 + c_1n_t + c_0.$$  \tag{119}

Values of the coefficients, as calculated by Perez, are given in Table 9. This model was also modified by Perez and Ineichen to forecast DNI from GHI forecasts provided by the operational model as well as corrections for locations presenting complex arid terrain [202].

### 6. Numerical Weather Prediction (NWP)

As discussed in the introduction of Section 5, the sparse nature of ground based pyranometer measurement stations limits the coverage and applicability of traditional time-series based forecasting models. Numerical Weather Predictions (NWPs) provide yet another alternative to a national or global scale ground based monitoring network. The development of NWPs began in the early 20th century with a paper written by Norwegian scientist Vilhelm Bjerknes titled, “The Problem of Weather Forecasting, Considered from the Viewpoints of Mechanics and Physics” [203]. Bjerknes knew that successful forecasting relied on sufficiently accurate knowledge regarding the state of the atmosphere at some initial time and the physical laws which governed the transition of the atmosphere from one state to another. This argument is still valid today and is at the heart of the NWP methodology. First a physical domain is chosen, then discretized at a desired resolution, and

### Table 8

Top: Regression coefficients for the Tarpley model as determined by Tarpley [193] and Raphael and Hay [180], in parenthesis. Bottom: Regression coefficients for the Justus-Paris-Tarpley model as determined by Justus et al. [194].

<table>
<thead>
<tr>
<th>Model</th>
<th>$a$</th>
<th>$b$</th>
<th>$c$</th>
<th>$d$</th>
<th>$e$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tarpley ($n &lt; 0.4$)</td>
<td>809.54 (-195.67)</td>
<td>3646.91 (3722.93)</td>
<td>1155.10 (85.98)</td>
<td>438.90 (151.10)</td>
<td>266.78 (-90.86)</td>
</tr>
<tr>
<td>Tarpley ($0.4 \leq n &lt; 1$)</td>
<td>800.79 (-199.30)</td>
<td>3959.34 (4047.97)</td>
<td>319.13 (-329.30)</td>
<td>(--)</td>
<td>(--)</td>
</tr>
<tr>
<td>Tarpley ($n = 1$)</td>
<td>274.73 (-49.80)</td>
<td>3672.04 (2178.16)</td>
<td>314.10 (-168.80)</td>
<td>(--)</td>
<td>(--)</td>
</tr>
<tr>
<td>Justus-Paris-Tarpley</td>
<td>0.4147</td>
<td>0.7165</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 9

Regression coefficients for the Perez Operational Satellite Model [16].

<table>
<thead>
<tr>
<th>$a$</th>
<th>$b$</th>
<th>$c_0$</th>
<th>$c_1$</th>
<th>$c_2$</th>
<th>$c_3$</th>
<th>$c_4$</th>
<th>$c_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0001</td>
<td>0.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>-0.58</td>
<td>-2.36</td>
<td>6.22</td>
<td>-6.2</td>
<td>2.36</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
physical laws of motion and thermodynamics are then numerically solved on the discrete spatial grid (see Fig. 15). It is the time marching of solutions to the physical and thermodynamic equations which provide the basis of NWP forecasts. However, Bjerknes recognized that the difficulty stems from the need to numerically solve a system of non-linear partial differential equations.

Four years after the work of Bjerknes, Austrian meteorologist Felix M. Exner produced reasonably accurate forecasts using an extremely simplified model based on the assumption of geostrophically balanced flow. Exner summarized his results in a textbook published in 1912 [204].

Soon after the work of Bjerknes and Exner, Lewis F. Richardson attempted to numerically solve weather forecast equations between 1916 and 1922 [205]. Richardson, as a proof of concept, computed a 6 h barotropic forecast by hand [205]. However, the method proved to be unsuccessful, resulting in a prediction two orders of magnitude too large. These errors were attributed to the imbalances in the initial data that were chosen [206–208]. In addition, the hand calculations themselves took longer than the forecast horizon of 6 h. Still, in 1922, Richardson published his findings and noted that $32 \times 2000 = 64,000$ computers would be needed to carry out computations faster than the weather advanced [205].

In 1928, Courant, Friedrichs and Lewy (CFL) studied the application of finite differences to weather forecasting equations and formulated the CFL condition which imposed a restriction on appropriate time steps given a spatial discretization [209,210]. The CFL condition, from a physical point of view, suggests that the maximum allowable time step must be on the same order of magnitude as the relevant time scales for the physical process in order to ensure stability in the integrations of the physical model; a necessary condition. When considering advection, for example, the maximum time scale is given by the ratio of the minimum distance between grid-points divided by the maximum velocity in the flow field [206]. As a result, as the spatial resolution increases, the time scale of the physical process and the maximum allowable time step decrease [211,212].

Furthermore, the systematic simplification of governing equations into a consistent set was originally facilitated through the development of filtered models by J. G. Charney and the quasi-geostrophic assumption [213]. Charney’s work, in conjunction with the work of A. Eliassen, improved the understanding of atmospheric motions and allowed for some of the first successful computer assisted numerical weather forecasts in the 1950s [214]. These forecasts were based on the Barotropic Vorticity Equation (BVE), given by

$$\frac{\partial \zeta_{b}}{\partial t} + \nabla_{h} \cdot \left( \zeta_{b} + f \right) \mathbf{v}_{b} = 0 \quad (120)$$

where $\zeta_{b}$ is the barotropic vorticity, $\nabla_{h}$ is the two-dimensional horizontal gradient, $\nabla_{h}^{2}$ is the two-dimensional horizontal Laplacian, $\mathbf{v}_{b}$ is the purely rotational two-dimensional wind velocity, $\psi_{b}$ is the barotropic streamfunction, $k$ is the unit vector normal to geopotential surfaces, and $f$ is the Coriolis parameter defined by (120).

$$f = 2\Omega \sin \phi. \quad (123)$$

The use of the BVE and the Electronic Numerical Integrator and Computer (ENIAC) at Aberdeen, Maryland in 1950 is widely regarded as the first computer assisted NWP model [206,215].

Soon after the successful application of the BVE at Aberdeen, NWP development began to grow in popularity with improved understanding of baroclinic instability through the work of J. G. Charney [216], E. T. Eady [217] and N. A. Phillips [218], refinement of the primitive equations in the work of J. Smagorinsky [219] and K. Hinkelmann [220,221], establishment of general circulation modeling by N. A. Phillips [222] and J. M. Lewis [223], and a great body of work on spectral methods and the transform method [224–228], semi-implicit time differencing [229–231] and atmospheric initialization [232–237]. What followed this early NWP was a vast number of progressively sophisticated approaches. These developments, facilitated through escalating computational power, allowed for increasingly higher resolutions and larger domains to be covered. As their scope is too broad to be covered here see Refs. [205,206,238,239] for more information.

Modern NWP models may be divided into two main categories: global or regional, depending on the domain which is used (see Fig. 16). At the current stage in their development, NWPs are unable to predict the precise position and extent of cloud fields. Their relatively coarse spatial resolution, typically on the order of 1–20 km (see Table 10) renders NWP models unable to resolve the micro-scale physics that are associated with cloud formation. As a result, cloud prediction inaccuracy is among the largest sources of error in an NWP based solar forecast. Therefore, examination of fields explicitly predicted by the NWP are used for the diagnoses of atmospheric features such as clouds and precipitation. For example, saturation (or near-saturation) of the atmosphere is typically required for the formation of clouds. During post-processing of the NWPs’ solution field, locations near atmospheric saturation are associated with cloud formation. This results in a cloud formation probability at best, which is inherently problematic for solar forecasts.

Fig. 15. Basis of all Numerical Weather Prediction (NWP) models. First a domain is defined. Secondly, the domain is spatially discretized at a desired resolution. Finally, the NWP predicts desired information by solving equations of motion and thermodynamics. The forecast shown is an 18 h ahead prediction of surface temperature by the NOAA RAP model using the CONUS domain on June 1, 2012.
6.1. Global Forecast System (GFS)

One of the most well known global NWP models is the Global Forecast System (GFS). The GFS model is run by NOAA every 6 h and produces forecasts out to 384 h (16 days) in advance on a 28 km × 28 km grid for the global domain [242]. The GFS loop time steps are 6 h up to 180 h (7.5 days), then change to 12-h time steps out to 384 h (16 days). In addition to the 28 km × 28 km horizontal discretization, the GFS models 64 vertical layers of the atmosphere. The RTM of the GFS accepts as inputs predicted values of a fully three-dimensional aerosol concentration. Aerosols are not explicitly considered except for a top of the atmosphere model. The GFS model was created by the National Center for Atmospheric Research (NCAR) in 2004. The WRF has, since its release, been shown that NWPs provide more accurate forecasts than satellite-based methods for time horizons exceeding 4 h [240,241]. Accordingly, NWPs provide the most attractive option for medium-to-long-term atmospheric forecasting.

6.2. Regional NWP models

Unlike global NWP models, regional NWP models simulate only a sub-domain of the global space (see Fig. 16). Regional models in the U.S. include the Rapid Update Cycle (RUC), RAPid refresh (RAP), North American Mesoscale (NAM) model, High Resolution Rapid Refresh (HRRR) and the Weather Research and Forecasting (WRF) models. Details of each are discussed in the following sections.

6.2.1. Rapid Update Cycle (RUC)/RAPid refresh (RAP) models

The RUC was a NOAA/NCEP (National Centers for Environmental Prediction) operational NWP model until May, 2012. RUC produced hourly updated 13 km × 13 km horizontally resolved forecasts with 50 atmospheric layers out to a time horizon of 18 h. The RUC loop time steps are 1 h from time of analysis out to 18 h. The RUC possessed a wavelength independent model for the absorption/scattering of radiation by water vapor only. Other atmospheric gases and aerosols were neglected. The RUC also assumes a simplified Rayleigh scattering model which failed to capture the complex relationship between scattering intensity and radiation wavelength. In addition, only downwelling irradiance was attenuated which sometimes leads to the underestimation of diffuse irradiance due to backscattering [244].

As of May 1, 2012 the RUC was replaced with the RAPid refresh (RAP) model as the next-generation version of the NCEP hourly cycle system. The RAP model possesses the same spatial and atmospheric resolution (12 km × 12 km, 50 layers) but is based on a new rapid update configuration of the WRF model (see Section 6.2.4). As a result, the RAP benefited from the ongoing community improvements to the WRF. The domain of the RAP is also significantly larger than the previous RUC and was expanded from the CONtinental United States (CONUS) region to include Alaska as well (see Fig. 16).

6.2.2. North American Mesoscale (NAM) model

The North American Mesoscale (NAM) model is the NCEP’s primary mesoscale environmental modeling tool. NAM produces 12 km × 12 km horizontally resolved forecasts with 60 atmospheric layers out to a time horizon of 96 h over North America and is updated four times daily. The NAM model loop time steps are 6 h from the time of analysis out to 84 h (3.5 days). The NAM model uses predicted water vapor concentrations, seasonally varying but zonally constant O₃ concentrations and constant CO₂ concentrations. Aerosols are not explicitly considered except for a top of the atmosphere adjustment, which is not particularly troublesome with the exception of regions with high levels of time varying aerosol concentrations. Wavelength specific attenuation of both upwelling and downwelling fluxes is accounted for.

6.2.3. High Resolution Rapid Refresh (HRRR) model

The High Resolution Rapid Refresh (HRRR) model is an NOAA operated, experimental, hourly updated, 3 km × 3 km resolution atmospheric model. The HRRR was previously only nested over the eastern 2/3 of the continental United States, but as of June 2009 coverage was expanded to the CONUS region similar to the former RUC (see Fig. 16). The HRRR model uses the 13 km resolution RUC/RAP for its initial conditions and is updated hourly. Benefits of the HRRR include the increased resolution and frequent updates which allow for shorter timescale predictions.

6.2.4. Weather Research and Forecasting (WRF) model

Some of the NWPs discussed were based on a version of the Weather Research and Forecasting (WRF) model which was created through a partnership between NOAA and the National Center for Atmospheric Research (NCAR) in 2004. The WRF has, since its
introduction, seen increased applicability in both research and operational communities. WRF software is supported by ongoing efforts including workshops and on-line documentation. One of the main goals of the WRF model is to advance mesoscale atmospheric prediction by promoting closer ties between research and operational forecasting communities. The WRF is flexible by design and intended for a wide variety of forecasting applications with a priority on spatial resolutions ranging from 1 to 10 km.

7. Local sensing

7.1. Sky imagers

Both NWP and satellite imaging techniques lack the spatial and temporal resolution to provide information regarding high frequency fluctuations of solar irradiance. An alternative is provided through ground based imaging of local meteorological conditions. An instrument used to image the hemispherical sky is called a “sky imager” or a “total sky imager”. An example of a sky imager is the Total Sky Imager (TSI) manufactured by Yankee Environmental Systems shown in Fig. 17, and an example of a TSI image taken for the Solar Forecasting Initiative at the University of California Merced is shown in Fig. 18. This model was not developed for solar forecasting applications and has some limitations in the image resolution and the presence of the shadow band that blocks a significant portion of the sky. Other devices such as the one developed at the Leibniz Institute of Marine Sciences at the University of Kiel [245] or the one developed by Cazorla et al. [246] were designed to overcome these limitations.

These sky imagers have been used to model irradiance at the ground level. For instance, Kalisch and Macke [245] used the images collected on four transatlantic research cruises to optimize parameterizations of solar downward radiation at the sea surface. Martínez-Chico et al. [247] used the sky imager to characterize cloud cover and prevailing sky situations in the south of Spain in order to assess the location’s suitability for sitting solar energy systems. Similarly Antón et al. [248] quantified the very short-term variability of the total solar irradiance in Granada, Spain and correlated the variability with the cloud cover obtained from a sky imager.

One of the key issues when using a sky imager is extracting cloud cover information from the images. Most of the cloud classification algorithms applied to sky images result into a binary mask (cloud/no-cloud) that cannot differentiate between thin and opaque clouds. Several researchers have tried to correct this problem. For example, Ghonima et al. [249] developed a classification algorithm that compares the pixels’ red to blue ratio against the values from a clear sky library. Additionally with the aid of the aerosol optical depth measured by an AERONET photometer they added a haze correction factor to the classification model to account for variations in the level of aerosol. Similarly, Calbó and Sábburg [250] used the sky images as inputs to an automatic classification algorithm that distinguishes between different types of sky conditions: clear, low cumuliform clouds, stratiform clouds, etc. The classification is based on features such as statistical measurements of image’s texture and its Fourier transform.

Besides using sky images to monitor and characterize sky conditions, researchers have used the sky images for the forecasting of solar irradiance. The methodology for processing ground based images is mostly similar to satellite based techniques. Projections of observed solar radiation conditions are based on immediate measured history while the position and impact of clouds is deduced from their motion. In the case of TSIs a CCD camera image is digitally processed in order to detect locations of the sky covered by clouds. The cloud image is then propagated forward in time resulting in a forecast. Chow et al. used images from a TSI to produce intra-hour forecasting of GHI of a distributed network of point sensors located at the University of California, San Diego [251]. Cloud fields were propagated forward in time using motion vectors generated by cross-correlating two consecutive images. The authors suggest that TSI images are useful for prediction of GHI on time horizons out to 15 min [251].
Crispim et al. [252] used AI techniques and cloudiness indices obtained from pixel classification of TSI images. These indices were used as inputs to ANN, optimized with a GA, for prediction of solar irradiance. Ferreira et al. [253] created their own portable sky imager and created ANN models optimized via a multi-objective genetic algorithm to predict GHI, cloudiness and temperature for forecasting horizons ranging from 5 min out to 4 h. Coimbra and Marquez also made use of a TSI for intra-hour DNI forecasts rather than GHI [254]. The work in Ref. [254] focused on detailed image processing procedures along with Particle Image Velocimetry (PIV) to determine grid cloud fractions in an area of interest. In agreement with the work of Chow et al. Coimbra and Marquez showed that TSIs show promise for forecasting horizons out to 15 min with the lowest error associated with time horizons of 5–6 min.

It should be noted that while these sky imaging-based methodologies make use of local meteorological information enabling intra-hour forecasts, their time horizon is restricted to approximately 30 min due to residence time of clouds over the field of view. One possible approach to extend the time horizon of ground based measurements is to distribute an array of imagers so that more information regarding local cloud fields is obtained. However, the relatively high cost associated with the TSI and the dynamic nature of local cloud fields, which may limit the correlation of successive images, pose difficulties for current ground based imaging methodologies. In addition to an upper bound on the time horizon of the TSI, a lower bound is also imposed. The lower bound is a result of circumsolar glare and limitations introduced by the shadow band, which renders time horizons shorter than 2 min inaccessible [251,254]. Nonetheless, sky imaging-based forecasts will continue to be relevant as a result of their unique temporal and spatial resolution.

7.2. Wireless sensor networks

None of the methods discussed so far possesses intra-minute time horizons. As discussed TSIs are limited to time horizons higher than 3 min, satellite and NWP models typically possess time horizons on the order of 30 min while stochastic and AI methods have not been widely applied to time horizons less than 15 min. Semiconductor point sensors are capable of very high sampling frequencies but fail to correctly characterize the dynamic nature of an operational scale PV plant [255]. An alternative has been suggested by A. Cerpa and coworkers at the University of California, Merced [256]. A 1 MW PV array was outfitted with 40 TelosB nodes equipped with low cost solar irradiance sensors (see Fig. 19(a)). The authors proposed a forecasting algorithm which used multiple readings from the spatially distributed network of sensors to compute future values of the distributed power output. The forecasting approach used spatial cross-correlations between sensor nodes which provided forecasts in the range of 20–50 s. Calculated velocities agreed with TSI calculated cloud velocity field 70% of the time. This work demonstrates the potential of wireless sensor networks as low cost and highly accurate approaches for intra-minute solar forecasting.

7.3. Pyranometer arrays

Two other methods have been developed by J. L. Bosch et al. to estimate cloud speed using radiometric measurements from sensors located at the UC San Diego Solar Energy test bed [257]. While these methods have not been successfully employed as part of a functioning forecast, the approach shows promise for irradiance prediction on intra-minute time horizons. Both methods are based on time-series analysis of GHI measurements using photodiode pyranometers (Li-200Z, Licor, Inc.) logged to a single CR1000 (Campbell Scientific Inc.) datalogger at an acquisition frequency of 20 Hz. The sensors were arranged in a semicircular array with a radius of 6 m (see Fig. 19(b)).

The first method is based on the premise that GHI time-series for a pair of sensors aligned with the direction of cloud movement will be highly correlated. In addition, the largest correlation should occur when the upwind sensor time-series is lagged by the travel time of the cloud edge between the two sensors. As a result, once a time lag is determined from the most correlated pair, the direction of the cloud motion is given by the alignment of the sensors and their distance by

\[ v = \frac{D}{t_{MCP}} \]  

(124)

where \( v \) is the cloud edge velocity, \( D \) is the separation between the sensors and \( t_{MCP} \) is the time lag between the most correlated pair of sensors. The time lag is calculated through a maximization of cross-correlations between time-series.

The second method used only three of the sensors in Fig. 19(b): the sensors at the origin, 0° and 90°. Consequently, the simple Equation (124) fails because the cloud edge motion directions may not necessarily be aligned with the axis of the sensor setup. Consequently, the most correlated pair method cannot be applied to the three sensors configuration. Bosh et al. developed an alternative method based on the assumptions of a linear cloud edge passing over the array of sensors and a constant cloud edge velocity (see Ref. [257]).

These methods were validated using data from October 20–21, 2011 taken at UCSD. These days were characterized by a large number of cumulus and stratocumulus clouds which resulted in relatively large variability in the GHI time-series. High variability in
the GHI signal is necessary in order to identify the most correlated pair of sensors. It should be noted that overcast and clear periods produce time-series of relatively low variability, resulting in a high correlation across all sensors, rendering the most correlated pair approach non-operational [257]. Cloud speeds determined by the two methods were compared to radiosonde data from the Integrated Global Radiosonde Archive (IGRA) [258], including profiles of pressure, wind speed and direction, and relative humidity in addition to METAR surface weather data which were obtained from the NOAA database [259]. Both methods show good agreement with the radiosonde and METAR data. The method that uses the most correlated pair has the advantages of simplicity, robustness and computational speed while the linear cloud edge method benefits from reduced installation and maintenance costs. However, it is important to note that both methods require high variability in the GHI time-series in order to generate highly correlated pairs of data.

8. Hybrid systems

A Hybrid System (HS) is characterized by a combination of any two or more of the methods described previously. As of late, several HSs have been used to produce high quality solar irradiance forecasts. One of the motivations for the development of hybrid models is that often it is possible to increase the forecasting accuracy by taking advantage of the strengths of each methodology. A typical example for this is the pairing of cloud cover information derived from satellite images with ANNs as implemented in Ref. [260]. Cloud speeds determined by the two methods were compared to radiosonde data from the Integrated Global Radiosonde Archive (IGRA) [258], including profiles of pressure, wind speed and direction, and relative humidity in addition to METAR surface weather data which were obtained from the NOAA database [259]. Both methods show good agreement with the radiosonde and METAR data. The method that uses the most correlated pair has the advantages of simplicity, robustness and computational speed while the linear cloud edge method benefits from reduced installation and maintenance costs. However, it is important to note that both methods require high variability in the GHI time-series in order to generate highly correlated pairs of data.

Sfetsos and Coonick introduced a new approach for the forecasting of mean hourly GHI using traditional linear stochastic methods, several artificial-intelligence-based techniques alongside the adaptive neuro-fuzzy inference scheme [78]. Zarzalejo et al. used ANNs in conjunction with satellite-derived cloud indices for the forecasting of hourly mean GHI [199]. Mellit et al. developed a hybrid model which was used to predict the daily GHI by combining an ANN and a library of Markov transition matrices [81]. Chaabene and Ammar introduced a medium term dynamic forecasting model for irradiance and ambient temperature which consisted of a neuro-fuzzy estimator based on the meteorological parameters’ behaviors during the days before as well as a short term forecast for 5 min ahead based on stochastic models and Kalman filtering [77]. Reikard performed forecasting tests using regressions in logs, ARIMA, and unobserved components models, transfer functions, neural networks and hybrid models [89]. Reikard claimed that the best results were obtained using the ARIMA in logs, with time-varying coefficients. Martin et al. tested AR, ANNs and fuzzy-logic models for application to solar thermal power plants energy production planning [44]. Mellit et al. also developed an adaptive model for predicting hourly GHI and DNI using a dataset of measured air temperature, relative humidity, and direct, diffuse and global horizontal irradiance [261]. The adaptive model’s performance was compared against an MLP ANN. Ji and Chee [79] recently developed a new approach that contained two phases used to predict the hourly solar radiation series. In the detrending phase, several models are applied to remove the non-stationary trend lying in the solar radiation series. In the prediction phase, the ARMA model is used to predict the stationary residual series with a time delayed neural network. Voyant et al. [80] proposed a study of the contribution of exogenous meteorological data as time-series to their optimized MLP. To do so they compared different forecasting methods including a persistence forecast, ARIMA reference predictor, an ANN with preprocessing using only endogenous inputs and an ANN with preprocessing using both endogenous and exogenous inputs.

More recently, Marquez and Coimbra [262] developed and validated a medium-term solar irradiance forecast for both GHI and DNI based on stochastic learning methods, ground experiments and the NWS database. A genetic algorithm was used as the input selector for the ANN in order to select the most relevant input patterns. Voyant et al. used a hybrid ARMA/ANN with NWPs to predict hourly mean GHI [263]. Marquez et al. [264] also used cloud indices obtained from a TSI and cloud indices derived from infrared radiometric measurements in order to improve results for hourly forecasts of GHI. Marquez, Pedro and Coimbra have developed a

![Fig. 20. Comparison of time horizon and spatial resolution. Solid lines indicate current limits of techniques while the dashed lines and arrows indicate the future progress of work. At techniques will continue to include local, mesoscale and global meteorological data which will allow for both shorter time horizons and greater areas of interest. In addition, recent trends in NWP suggest that shorter time horizons will be available through the development of the WRF models.](image-url)
forecast based on satellite images and ANNs for time-series predictions of GHI out to 2 h ahead [260]. It is noteworthy in Fig. 20 that Stochastic and ANN techniques have a much broader coverage than other techniques especially in terms of time horizon. This is a direct result from these techniques relying on observational time-series and therefore their time horizon is generally limited only by the sampling frequency used for data acquisition. However, this does not imply that an appropriate ratio between time horizon and temporal resolution should not be conserved. In other words, as the sampling frequency changes, so should the time horizon. For instance, a sampling frequency of 2 Hz would imply the same bound on temporal resolution, therefore a forecast horizon on the order of seconds would be appropriate. Although, if high-frequency data is available it is possible to pre-process the data in order to produce, for example, hourly or daily averages, lowering sampling frequencies and allowing for an effective extension of the temporal horizon. Having said that, the reciprocal argument does not hold: data initially logged at low sampling frequencies, daily averages for instance, cannot be pre-processed in a manner which would make it appropriate for forecasting short time horizons; e.g. 30 min. In conclusion, meaningful time horizons should be of the same order of magnitude as the temporal resolution of the data. The degradation of forecasting quality with increased temporal extrapolation is a common issue in the area of forecasting and is not unique to irradiance prediction. In addition to the limitations imposed on spatial and temporal resolution, it should be noted that the forecasting skill of each of the individual techniques is a function of these resolutions; most notably the time horizon. Fig. 21 shows the performance metric for several techniques from six publications as a function of time horizon. As discussed in Ref. [27], the performance metric is very well approximated by the ratio of the model’s RMSE to the RMSE of the persistence model

\[ s \approx 1 - \frac{\text{RMSE}}{\text{RMSE}_p}. \]  

(125)

Here it is clear that the forecasting methods are not only range-bound but their forecasting skill is a function of time horizon. As information regarding the RMSE is available in several publications, this approximation of the performance metric was used in Fig. 21. Several methods are shown for time horizons ranging from 30 to 120 min. In this temporal range the hybrid GA/ANN from Ref. [148] performs the best.

9. Performance of forecasting techniques

In this section we review results from a selected group of recent papers that illustrate the state of the art in solar forecasting. The timeline for the works included in this comparison starts in 2008 (for articles published prior to 2008 we refer the reader to the comprehensive review by Mellit and Kalogirou [95]). Table 11 summarizes some of these works and lists their main features such as the type of data used, the forecasting horizon and the methodology used. In this section we present results compiled from some of these references grouped by the type of forecasting technique. However, sometimes it is difficult to say whether a paper belongs with one category or the other based on the methodology since most of the recent research in this field seldom relies on a single family of techniques. For those cases we used our best judgment based on the emphasis given to the different techniques employed in the articles. Wherever possible we computed the forecast skill as defined above in Eq. (125) in order to enable a quantitative comparison between the different techniques.

9.1. NWP-based forecast

Even though many NWP models are run daily and available in real-time for virtually any location on earth (after all, they are the basis for most of the day-to-day weather forecasts) there are only a few studies benchmarking NWP radiation forecast against ground truth data. Understandably these studies concentrate on locations in the USA, Spain and Germany as these countries are among the world leaders in installed photovoltaic power capacity.

One of the most thorough benchmarking studies was conducted by Perez et al. (2010) [43]. In that work, cloud cover forecasts from the US National Digital Forecast Database (NDFD) were used to forecast hourly averages of GHI 1–7 days ahead of time. For short-term forecasts, 1–6 h ahead of time, GHI was predicted from extrapolated cloud cover derived from consecutive satellite images. The forecasted values were compared against a year’s worth of ground measurements for seven locations in the United States. Mathiesen and Kleissl (2011) [265] used data from the same array of ground sensors to assess the forecasting performance of three NWP models – North American Model (NAM), Global Forecast System (GFS), and European Centre for Medium-Range Weather Forecasts (ECMWF). The models were used to predict intra-day (0–24 h) hourly averages of GHI. Additionally they applied a bias correction function to the NWP forecasts, obtained through Model Output
### Table 11
Summary of recent publications in the forecasting of solar irradiance, PV power output and cloud cover.

<table>
<thead>
<tr>
<th>Authors</th>
<th>Year</th>
<th>Ref.</th>
<th>Forecast variable</th>
<th>Forecast horizon</th>
<th>Method</th>
<th>Exogenous variables</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hocaoglu et al.</td>
<td>2008</td>
<td>[286]</td>
<td>Hourly GHI</td>
<td>1 h</td>
<td>ANNs and preprocessing (hourly GHI is rendered as a 2-D image)</td>
<td>–</td>
<td>1 year of hourly GHI from Iki Eylul, Turkey</td>
</tr>
<tr>
<td>Lorenz et al.</td>
<td>2009</td>
<td>[241]</td>
<td>Hourly GHI</td>
<td>1 h to 3 days</td>
<td>ECMWF and post-processing</td>
<td>–</td>
<td>11 months of GHI from 200 weather stations and PV data from 11 systems, Germany</td>
</tr>
<tr>
<td>Gordon</td>
<td>2009</td>
<td>[89]</td>
<td>5, 15, 30, and 60 min averaged GHI</td>
<td>5, 15, 30 min and 1–4 h</td>
<td>Regressions in logs, ARIMA, ANNs, and hybrid (ARIMA and ANN)</td>
<td>Humidity, cloud cover, atmospheric turbulence (when available)</td>
<td>Hourly GHI data for 4 years in 3 locations and 1 min GHI for 1 year in 3 locations, USA</td>
</tr>
<tr>
<td>Bacher et al.</td>
<td>2009</td>
<td>[270]</td>
<td>Hourly PV power</td>
<td>Out to 36 h</td>
<td>AR, ARX, RX (regressive model with no endogenous variables) and clear normalization</td>
<td>Forecasted GHI (HIRLAM mesoscale NWP from the Danish Meteorological Institute)</td>
<td>1 year of 15-min PV data from 21 rooftops systems, Denmark.</td>
</tr>
<tr>
<td>Perez et al.</td>
<td>2010</td>
<td>[43]</td>
<td>Hourly GHI</td>
<td>1–6 h, 1–7 days</td>
<td>Empirical fit between sky cover and measured GHI</td>
<td>Cloud cover forecasts</td>
<td>1 year of hourly GHI from the SURFRAD network, USA</td>
</tr>
<tr>
<td>Martin et al.</td>
<td>2010</td>
<td>[44]</td>
<td>Half daily values of GHI</td>
<td>Out to 3 days</td>
<td>AR, ANNs and fuzzy logic</td>
<td>–</td>
<td>Several years of hourly GHI from the Spanish National Weather Service, Spain</td>
</tr>
<tr>
<td>Mellit et al.</td>
<td>2010</td>
<td>[261]</td>
<td>Hourly GHI, DNI and DHI</td>
<td>1 h</td>
<td>ANNs and AR model (with parameters that are a non-linear combinations of exogenous variables)</td>
<td>Sunshine duration, temperature, humidity</td>
<td>5 years of 5 min GHI data for Jeddah, Saudi Arabia</td>
</tr>
<tr>
<td>Paoli et al.</td>
<td>2010</td>
<td>[138]</td>
<td>Daily GHI</td>
<td>1 day</td>
<td>kNN, ANNs, AR, Markov chains, Bayesian inference and clear sky normalization</td>
<td>–</td>
<td>19 years of daily GHI from Ajaccio, France</td>
</tr>
<tr>
<td>Mathiesen and Kleissl</td>
<td>2011</td>
<td>[265]</td>
<td>Hourly GHI</td>
<td>1 h to 1 day</td>
<td>NWP (NAM, GFS, and ECMWF) and bias correction</td>
<td>–</td>
<td>Hourly GHI from the SURFRAD network, USA</td>
</tr>
<tr>
<td>Marquez and Coimbra</td>
<td>2011</td>
<td>[262]</td>
<td>Hourly GHI and DNI</td>
<td>Out to 6 days</td>
<td>NWP and ANN and Gamma Test for input selection</td>
<td>NWP weather forecast</td>
<td>1 year of 30 s GHI and DNI data for Merced, USA</td>
</tr>
<tr>
<td>Chow et al.</td>
<td>2011</td>
<td>[251]</td>
<td>Cloud cover and GHI</td>
<td>30 s to 5 min</td>
<td>Advection of the 2D cloud map; cloud motion generated from image cross-correlation; GHI obtained from empirical relation with cloud cover</td>
<td>–</td>
<td>Several days of 1 s GHI and 30 s sky-images in San Diego, USA</td>
</tr>
<tr>
<td>Ji and Chee</td>
<td>2011</td>
<td>[79]</td>
<td>Hourly GHI</td>
<td>1 h</td>
<td>Hybrid (ANN and ARMA) and clear sky normalization</td>
<td>–</td>
<td>1 year of 10 min data from Nanyang, Singapore</td>
</tr>
<tr>
<td>Voyant et al.</td>
<td>2011</td>
<td>[80]</td>
<td>Daily GHI</td>
<td>1 day</td>
<td>ANN, AR, ARIMA; clear sky normalization</td>
<td>Pressure, nebulosity, humidity, wind speed, etc.</td>
<td>9 years of 30 min data from the French Meteorological Organization in Corsica, France</td>
</tr>
<tr>
<td>Lorenz et al.</td>
<td>2011</td>
<td>[291]</td>
<td>Hourly PV power</td>
<td>Out to 2 days</td>
<td>ECMWF and post-processing to forecast GHI; explicit physical modeling to convert GHI into PV power</td>
<td>GHI</td>
<td>10 months of regional, hourly PV power data for Germany</td>
</tr>
<tr>
<td>Pelland et al.</td>
<td>2011</td>
<td>[267]</td>
<td>Hourly GHI and PV power</td>
<td>Out to 2 days</td>
<td>Canadas GEM and post-processing to forecast GHI; explicit physical modeling to convert GHI into PV power</td>
<td>Temperature</td>
<td>2 years of data: GHI obtained from 10 North-American ground stations, and PV power from 3 PV systems located in Canada</td>
</tr>
<tr>
<td>Capizzi et al.</td>
<td>2012</td>
<td>[287]</td>
<td>Daily GHI</td>
<td>1 day</td>
<td>Wavelet recurrent neural networks</td>
<td>Wind speed, humidity, and temperature</td>
<td>1 year of GHI and meteorological data in Catania, Italy</td>
</tr>
<tr>
<td>Chow et al.</td>
<td>2012</td>
<td>[288]</td>
<td>10 min PV power</td>
<td>10 and 20 min</td>
<td>ANN</td>
<td>Solar geometry, temperature, irradiance</td>
<td>2 weeks of 10 min PV data and hourly meteorological data for Hong Kong, China</td>
</tr>
<tr>
<td>Yap and Karri</td>
<td>2012</td>
<td>[289]</td>
<td>Monthly GHI</td>
<td>1 month</td>
<td>Linear regression, AngstromPrescottPage and ANNs</td>
<td>Temperature, rainfall, evaporation, sunshine hours</td>
<td>12 years of daily GHI and meteorological data for Darwin, Australia</td>
</tr>
<tr>
<td>Lara-Fanego et al.</td>
<td>2012</td>
<td>[269]</td>
<td>Hourly GHI and DNI</td>
<td>1, 2 and 3 days</td>
<td>GHI: direct from WRF model; DNI: WRF model outputs and satellite measurements of aerosol and ozone levels</td>
<td>–</td>
<td>2 years of hourly GHI from 4 ground stations, Spain</td>
</tr>
<tr>
<td>Mandal et al.</td>
<td>2012</td>
<td>[273]</td>
<td>Hourly PV power</td>
<td>1 h</td>
<td>ANN and wavelet decomposition</td>
<td>GHI, Temperature</td>
<td>1 year of hourly GHI, temperature and PV power from Oregon, USA NREL's third typical meteorological year (TMY3) for Orlando and Miami, USA</td>
</tr>
<tr>
<td>Yang et al.</td>
<td>2012</td>
<td>[271]</td>
<td>Hourly GHI</td>
<td>1 h</td>
<td>ARIMA and seasonal decomposition</td>
<td>DNI, DHI, cloud cover indices</td>
<td></td>
</tr>
</tbody>
</table>
Table 11 (continued)

<table>
<thead>
<tr>
<th>Authors</th>
<th>Year</th>
<th>Ref.</th>
<th>Forecast variable</th>
<th>Forecast horizon</th>
<th>Method</th>
<th>Exogenous variables</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pedro and Coimbra</td>
<td>2012</td>
<td>[148]</td>
<td>Hourly PV power</td>
<td>1 and 2 h</td>
<td>ARIMA, ANNs, kNN, clear sky normalization and GA optimization</td>
<td>–</td>
<td>1 year of hourly PV power from a 1 MW solar farm in Merced, USA</td>
</tr>
<tr>
<td>Voyant et al.</td>
<td>2012</td>
<td>[263]</td>
<td>Hourly GHI</td>
<td>1 h</td>
<td>ARMA and ANNs</td>
<td>NWP weather forecast</td>
<td>6 years of hourly data from 5 locations in the Mediterranean, France</td>
</tr>
<tr>
<td>Marquez et al.</td>
<td>2013</td>
<td>[260]</td>
<td>30, 60, 90, and 120 min averaged GHI</td>
<td>30, 60, 90, and 120 min</td>
<td>ANNs and satellite images</td>
<td>–</td>
<td>1 year of 1 min GHI data for Davis and Merced and hourly NOAA’s GOES West satellite images, USA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marquez and Coimbra</td>
<td>2013</td>
<td>[264]</td>
<td>Hourly GHI</td>
<td>1 h</td>
<td>ANN and 3 cloud cover indices (visible radiation, sky-images and infrared radiation)</td>
<td>Cloud cover, IR radiation, sky-images</td>
<td>Several days of 1 min GHI and 20 s sky-images for Merced, USA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bosch et al.</td>
<td>2013</td>
<td>[257]</td>
<td>Cloud velocity</td>
<td>–</td>
<td>Estimates cloud speed using measurements from 8 GHI sensors</td>
<td>GHI</td>
<td>Several days of very high frequency GHI data for San Diego, USA</td>
</tr>
<tr>
<td>Voyant et al.</td>
<td>2013</td>
<td>[290]</td>
<td>Hourly GHI</td>
<td>1 h</td>
<td>Hybrid (ANN and ARMA); season dependent model; seasonal ANN learning; lag-error dependent model; clear sky normalization</td>
<td>Nebulosity, pressure, precipitation</td>
<td>10 years of hourly GHI in 5 locations in the Mediterranean, France</td>
</tr>
<tr>
<td>Yona et al.</td>
<td>2013</td>
<td>[292]</td>
<td>Hourly PV power</td>
<td>1 h</td>
<td>Fuzzy theory to predict GHI from NWP weather data; recurrent NN TSK fuzzy models</td>
<td>Cloud cover, humidity, temperature</td>
<td>1 year of data for Okinawa, Japan</td>
</tr>
<tr>
<td>Jafarzadeh et al.</td>
<td>2013</td>
<td>[293]</td>
<td>15 min PV power</td>
<td>15–60 min</td>
<td></td>
<td>Temperature, cloud cover</td>
<td>1 min data from a power plant in Reno, USA</td>
</tr>
</tbody>
</table>

Statistics (MOS), in order to minimize the forecast’s bias error. Tables 12 and 13 list the forecasting performance obtained in these two studies. Direct comparison is not possible given that the reported RMSEs in Mathiesen and Kleissl (2011) are not separated by forecasting horizon as in the case of Perez et al. (2010).

One of the main conclusions from these benchmarking studies was that most operational NWP models over-predict irradiance. In a recent study Mathiesen et al. [266] analyzed the predictions from NAM and concluded that this model predicts insufficient cloud cover on the California coast leading to the said irradiance over-prediction. In this work the authors presented and tested some techniques in order to obtain a more realistic cloud cover prediction, namely the improvement of the NAM’s initial condition by assimilating realistic could cover retrieved from satellite images, and the optimization of the model’s resolution and physics parameterizations. These measures largely mitigate the bias observed in NAM’s GHI forecasts and improved the forecasting accuracy both for intra-day and day ahead time horizons.

Motivated by the need to predict rooftop PV power in Germany, Lorenz et al. [241] computed hourly GHI from three-hourly ECMWF forecasts out to 3 days ahead of time. The predicted GHI was then used to forecast PV power output. The forecasting performance for the time horizon of 1 day was computed using data from 200 weather stations that cover the entire German territory. With so much ground data available the authors were able to study the forecasting performance as a function of spatial averaging by assembling forecasts for distributed stations. That analysis was one of the most interesting and original aspects of that article as rarely researchers have access to such a wealth of data. Similarly, Pelland et al. [267] used the Environment Canadas Global Environmental Multiscale NWP model to forecast hourly GHI and PV power for horizons out to 48 h. They applied spatial averaging and bias removal using a Kalman filter to the NWP forecasts in order to increase the predictions’ accuracy.

In these studies the researchers could simply download forecasted GHI data from the several services that provide NWP forecasts. That is not possible if one needs to predict DNI or DHI as they are rarely part of the NWP output variables. To fill this gap Breitkreuz et al. [268] proposed a model based on information provided by air-quality models to enable NWP models to compute DNI irradiance. Alternatively, Lara-Fanego et al. (2012) [269] took advantage of the fact that the WRF offers a detailed picture of the future state of the atmosphere to implement a physical post-processing technique to forecast DNI based on the WRF model outputs and satellite measurements of aerosol and ozone levels. In that work, hourly GHI and DNI forecasted with the WRF model

<table>
<thead>
<tr>
<th>Horizon</th>
<th>Satellite</th>
<th>NDFD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 h</td>
<td>2 h</td>
</tr>
<tr>
<td>RMSE [W/m²]</td>
<td>87.57</td>
<td>102.14</td>
</tr>
<tr>
<td>Forecasting skill, s</td>
<td>0.08</td>
<td>0.18</td>
</tr>
</tbody>
</table>

Table 12
Forecast skill for hourly averaged GHI vs. forecasting horizon. Table adapted from results in Ref. [43].

<table>
<thead>
<tr>
<th>Horizon</th>
<th>Satellite</th>
<th>NDFD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3 h</td>
<td>4 h</td>
</tr>
<tr>
<td>RMSE [W/m²]</td>
<td>116.29</td>
<td>127.14</td>
</tr>
<tr>
<td>Forecasting skill, s</td>
<td>0.17</td>
<td>0.17</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Horizon</th>
<th>Satellite</th>
<th>NDFD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5 h</td>
<td>6 h</td>
</tr>
<tr>
<td>RMSE [W/m²]</td>
<td>135.71</td>
<td>159.43</td>
</tr>
<tr>
<td>Forecasting skill, s</td>
<td>0.11</td>
<td>0.06</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Horizon</th>
<th>Satellite</th>
<th>NDFD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2 days</td>
<td>3 days</td>
</tr>
<tr>
<td>RMSE [W/m²]</td>
<td>157.86</td>
<td>164.86</td>
</tr>
<tr>
<td>Forecasting skill, s</td>
<td>0.18</td>
<td>0.21</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Horizon</th>
<th>Satellite</th>
<th>NDFD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4 days</td>
<td>5 days</td>
</tr>
<tr>
<td>RMSE [W/m²]</td>
<td>168.86</td>
<td>174.00</td>
</tr>
<tr>
<td>Forecasting skill, s</td>
<td>0.21</td>
<td>0.23</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Horizon</th>
<th>Satellite</th>
<th>NDFD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6 days</td>
<td>7 days</td>
</tr>
<tr>
<td>RMSE [W/m²]</td>
<td>180.57</td>
<td>191.43</td>
</tr>
<tr>
<td>Forecasting skill, s</td>
<td>0.21</td>
<td>0.14</td>
</tr>
</tbody>
</table>
were compared against ground measurements for several locations in southern Spain. The comparison was carried for forecasting horizons of 1, 2 and 3 days. Table 14 lists the average RMSE found in that work and the respective forecast skill. Similarly, Marquez and Coimbra (2011) [262] used NWP forecasts (e.g. sky cover, temperature, probability of precipitation, etc.) from the National Weather Service’s (NWS) database as exogenous inputs for ANNs to predict hourly GHI and DNI out to 6 days ahead of time for Merced California. The optimal set of input variables was determined via a genetic algorithm. Table 15 shows the RMSE for the same-day forecast obtained with the best model. The authors demonstrated that hybrid models which combine data generated by NWP algorithms and machine learning algorithms can yield even better results than just post-processing the NWP forecast.

All these studies are invaluable to assess the accuracy when using NWP data in the forecasting of solar irradiation. These works also show that clever post-processing techniques, such as MOS bias correction, the improvement the NWP models’ initial conditions and physical parameters or the use of machine learning techniques on top of the NWP models’ outputs can improve the forecasting accuracy substantially. Furthermore, they demonstrated that even when no irradiance data is available from NWP models the other weather variables can be very useful as exogenous inputs for irradiance forecasting models. Finally, a common remark present in these articles is that the lack of temporal resolution in the NWP forecasts makes them useless in predicting the intra-hour fast ramp rates observed in solar irradiance.

9.2. Stochastic forecasts

Bacher et al. (2009) [270] tested several stochastic models to predict hourly values of PV power for horizons out to 36 h. The authors used 15 min averaged data from 21 PV rooftops systems in Denmark. The PV data was normalized using a smooth curve fit for the clear sky PV output, and forecasts of the normalized solar power were calculated using AR and ARX models. The ARX models used forecasts of GHI provided by the Danish Meteorological Institute using the HIRLAM mesoscale NWP model as an exogenous variable. The authors proposed a third model (denoted as LMnwp) which is a special case of the ARX model in the way that it includes no lagged values of the endogenous variable (normalized PV time-series). The short- and medium-term forecast skills for the 3 models are listed in Table 16. The authors concluded that for short-term forecast (out to 2 h) the most important input is the lagged PV values, while for longer horizons the NWP GHI becomes more important.

Yang et al. (2012) [271] constructed three ARIMA forecasting models for next-hour GHI including cloud cover effects. The main difference in the three models tested concerns the inputs used: GHI in the first model, DNI and DHI in the second model and cloud cover (CC) in the third model. In the first two models the irradiance inputs were treated with an additive seasonal decomposition in order to account for deterministic irradiance variations. The authors used the third typical meteorological year (TMY3) data from the National Solar Radiation Data Base [272] to estimate the ARIMA models and to validate the forecasting accuracy.

9.3. AI forecasting models

Gordon (2009) [89] ran forecasting experiments for GHI using several data sets, at resolutions of 5, 15, 30, and 60 min. Forecasting tests were run using regressions in logs, ARIMA, and neural networks among other models to produce GHI forecasts 5, 15, 30 min and 1–4 h ahead of time. Hybrid models that combine ARIMA and ANN were also evaluated. The forecasting skill for the different models and forecasting horizons are listed in Table 17. The forecast skill is in reference to a simple forecasting model, which is a regression in logs. The authors argue that taking logs significantly reduces the forecast error at the transition points between sunrise and sunset.

Paoi et al. (2010) [138] used several stochastic and machine learning techniques to predict daily averaged global solar radiation one day ahead of time. Additionally, the authors applied two pre-processing strategies in order to normalize solar radiation. The one first was based on the clearness index and the second one used the clear sky index. Forecasting skills for some of the models developed by Paoi et al. (2010) are listed in Table 18.

Martin et al. (2010) [44] used AR, ANNs and fuzzy logic models to predict half-day values of GHI (averaged GHI from sunrise to noon and form noon to sunset) out to 3 days ahead of time. The forecasting accuracy was determined with ground solar radiation from the Spanish National Weather Service. Forecasting skill versus the time horizon for the different models is shown in Table 19.

Mandal et al. (2012) [273] used an ANN with exogenous variables to forecast the hourly PV power for forecasting horizon of 12 h. In order to improve the forecasting performance the authors applied a wavelet decomposition to the PV time-series, which is then input into the ANN, whose outputs are then passed through a wavelet reconstruction to determine the PV predictions. The substantial improvement obtained with this technique is reported in Table 20.

### Table 13
RMSE for NWP forecasts before and after MOS correction. Results averaged for the 7 sites analyzed [265].

<table>
<thead>
<tr>
<th>NWP model</th>
<th>RMSE [W/m²]</th>
<th>MOS-corrected RMSE [W/m²]</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAM</td>
<td>134.2</td>
<td>114.1</td>
</tr>
<tr>
<td>GFS</td>
<td>110.5</td>
<td>84.6</td>
</tr>
<tr>
<td>ECMWF</td>
<td>123.2</td>
<td>106.2</td>
</tr>
</tbody>
</table>

### Table 15
Same-day forecast performance for the best ANN models obtained in Ref. [262].

<table>
<thead>
<tr>
<th>Persistence</th>
<th>Best model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forecasting skill, s</td>
<td></td>
</tr>
<tr>
<td>GHI</td>
<td>123.10</td>
</tr>
<tr>
<td>DNI</td>
<td>270.00</td>
</tr>
</tbody>
</table>

### Table 14
WRF’s GHI and DNI forecast evaluation results as a function of the forecasting horizon for two locations in southern Spain from Ref. [269].

<table>
<thead>
<tr>
<th>Horizon [days]</th>
<th>Cordoba</th>
<th>Andasol</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE [W/m²]</td>
<td>s</td>
</tr>
<tr>
<td>WRF Persist.</td>
<td>WRF Persist.</td>
<td></td>
</tr>
<tr>
<td>GHI 1</td>
<td>170.0</td>
<td>0.11</td>
</tr>
<tr>
<td>2</td>
<td>175.0</td>
<td>0.13</td>
</tr>
<tr>
<td>3</td>
<td>185.0</td>
<td>0.05</td>
</tr>
<tr>
<td>DNI 1</td>
<td>304.0</td>
<td>0.16</td>
</tr>
<tr>
<td>2</td>
<td>311.0</td>
<td>0.23</td>
</tr>
<tr>
<td>3</td>
<td>319.0</td>
<td>0.21</td>
</tr>
</tbody>
</table>

### Table 16
Forecast skill for models proposed in Ref. [270].

<table>
<thead>
<tr>
<th>Forecasting skill, s</th>
<th>1–6 h</th>
<th>19–29 h</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR</td>
<td>0.27</td>
<td>0.17</td>
</tr>
<tr>
<td>LMnwp</td>
<td>0.25</td>
<td>0.36</td>
</tr>
<tr>
<td>ARX</td>
<td>0.35</td>
<td>0.36</td>
</tr>
</tbody>
</table>
Hammer et al. (1999, 2003) to extrapolate the temporal ANNs. Their work is an extension to the methods proposed by that combine information from processed satellite images with and UC Merced in California 120 min. The tests were carried out with data gathered at UC Davis developed to forecast GHI at temporal horizons of 30, 60, 90, and 9.4. Hybrid forecasting models function of the forecast horizon.

Table 21 lists the forecasting skills for the models developed as a solution of several problems simultaneously [276]. Researchers, renewable generation on the electrical power grids requires the industry experts and operators have identified a number of major challenges including: 1) unpredictable and steep ramps, 2) making up for errors in the forecasting of renewable resources, 3) intra-hour variability, and 4) over generation in the middle of the night. The first three issues are closely connected to the ability to forecast the resource and power plant output accurately. In fact, a perfect forecast would eliminate these problems entirely. Given that the perfect forecast is impossible to achieve, two solutions to mitigate these problems are currently under investigation: storage and demand response [277, 278].

Table 17
Forecast skill for several models as function of location and forecast horizon from Ref. [89].

<table>
<thead>
<tr>
<th>Forecast horizon</th>
<th>Clark, Nevada</th>
<th>National wind database</th>
<th>SRRL database</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ARMA</td>
<td>NN</td>
<td>Hybrid</td>
</tr>
<tr>
<td>5 min</td>
<td>0.19</td>
<td>0.13</td>
<td>0.19</td>
</tr>
<tr>
<td>15 min</td>
<td>0.28</td>
<td>0.21</td>
<td>0.28</td>
</tr>
<tr>
<td>30 min</td>
<td>0.39</td>
<td>0.28</td>
<td>0.39</td>
</tr>
<tr>
<td>1 h</td>
<td>0.29</td>
<td>0.23</td>
<td>0.29</td>
</tr>
<tr>
<td>2 h</td>
<td>0.16</td>
<td>0.12</td>
<td>0.16</td>
</tr>
<tr>
<td>3 h</td>
<td>0.13</td>
<td>0.12</td>
<td>0.12</td>
</tr>
<tr>
<td>4 h</td>
<td>0.085</td>
<td>0.053</td>
<td>0.085</td>
</tr>
</tbody>
</table>

Table 18
Forecast skill for some of the models in Ref. [138].

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE [MJ/m²]</th>
<th>Forecasting skill, s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persistence</td>
<td>4.65</td>
<td>–</td>
</tr>
<tr>
<td>kNN</td>
<td>4.48</td>
<td>0.04</td>
</tr>
<tr>
<td>AR</td>
<td>3.77</td>
<td>0.19</td>
</tr>
<tr>
<td>ANN</td>
<td>3.73</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Table 19
Forecast skill for several model from Ref. [44].

<table>
<thead>
<tr>
<th>Horizon [hours]</th>
<th>AR</th>
<th>ANN</th>
<th>ANFIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–12</td>
<td>0.12</td>
<td>0.12</td>
<td>0.11</td>
</tr>
<tr>
<td>12–24</td>
<td>0.16</td>
<td>0.17</td>
<td>0.17</td>
</tr>
<tr>
<td>24–36</td>
<td>0.18</td>
<td>0.20</td>
<td>0.19</td>
</tr>
<tr>
<td>36–48</td>
<td>0.19</td>
<td>0.21</td>
<td>0.20</td>
</tr>
<tr>
<td>48–60</td>
<td>0.20</td>
<td>0.23</td>
<td>0.21</td>
</tr>
<tr>
<td>60–72</td>
<td>0.21</td>
<td>0.23</td>
<td>0.22</td>
</tr>
</tbody>
</table>

Table 20
Forecast performance from models in Ref. [273].

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE [MJ/m²]</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPNN</td>
<td>2.07</td>
</tr>
<tr>
<td>RBFNN</td>
<td>1.35</td>
</tr>
<tr>
<td>WT + BPNN</td>
<td>1.35</td>
</tr>
<tr>
<td>WT + RBFNN</td>
<td>0.78</td>
</tr>
</tbody>
</table>

* Averaging the 8 separated days analyzed there.

Table 21
Forecast skill for the models employed in Ref. [148].

<table>
<thead>
<tr>
<th>Forecast</th>
<th>Persistence</th>
<th>RMSE [MW]</th>
<th>ARMA</th>
<th>kNN</th>
<th>ANN</th>
<th>ANFIS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>107.48</td>
<td>0.02</td>
<td>-0.08</td>
<td>0.18</td>
<td>0.31</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>160.79</td>
<td>0.10</td>
<td>-0.01</td>
<td>0.11</td>
<td>0.35</td>
<td></td>
</tr>
</tbody>
</table>

Pedro and Coimbra (2012) [148] applied several stochastic and AI techniques (ARIMA, kNN, ANN) to predict the 1- and 2-h averaged power output of a 1 MW solar power plant in Merced, California. In that study no exogenous variables were used. The authors created a clear sky model for the power output by smoothing the PV data. Additionally, a GA was employed to determine the ANNs’ architecture and input selection such that the forecasting performance was maximized. A detailed discussion of the seasonal dependency of the accuracy is presented in the article. Table 21 lists the forecasting skills for the models developed as a function of the forecast horizon.

9.4. Hybrid forecasting models

Marquez et al. (2013) [260] tested hybrid forecasting models that combine information from processed satellite images with ANNs. Their work is an extension to the methods proposed by Hammer et al. (1999, 2003) [274, 275] to extrapolate the temporal development of the clouds based on satellite images (obtained from the NOAA GOES West). The models in Marquez et al. (2013) were developed to forecast GHI at temporal horizons of 30, 60, 90, and 120 min. The tests were carried out with data gathered at UC Davis and UC Merced in California’s Central Valley. Table 22 lists the forecast skill for the persistence forecast and 2 ANN-based models. Input variables for Model 1 include only data derived form satellite images, while Model 2 also includes lagged GHI data in the set of input variables.

10. Other challenges of solar integration

As mentioned in the introduction, achieving higher levels of renewable generation on the electrical power grids requires the solution of several problems simultaneously [276]. Researchers, industry experts and operators have identified a number of major challenges including: 1) unpredictable and steep ramps, 2) making up for errors in the forecasting of renewable resources, 3) intra-hour variability, and 4) over generation in the middle of the night. The first three issues are closely connected to the ability to forecast the resource and power plant output accurately. In fact, a perfect forecast would eliminate these problems entirely. Given that the perfect forecast is impossible to achieve, two solutions to mitigate these problems are currently under investigation: storage and demand response [277, 278].

Energy storage is generally expensive, but offers the possibility of a definitive answer to the variability problem [279], but even in this solution accurate forecasting is still likely to be a critical component. Indeed, optimized storage systems and load-following as well as resource-following forecasts have been shown to increase the lifetime of batteries by reducing the number of deep discharging events. However, energy storage is a very broad field and covering this topic is beyond the scope of this review.

Demand response programs have been in place for several decades [280]. In general, demand response programs rely on the end-users deliberately altering the use of equipment and systems. In practice, this means changing lifestyle or reducing comfort, or changing operating procedures in exchange for some financial compensation or improved reliability of power supply. With the advent of the smart grid concept the industry is moving towards the concept of automated demand response (autoDR) where users are equipped with automated power management tools that respond to requests to decrease or increase the demand according to commands received from the systems operator. In this way, large consumers of energy are essentially energy reservoirs that can be tapped when increased load or diminished generation is expected elsewhere. Fig. 22 shows a block diagram for a type of automated demand response system. The ISO demand response automated...
server (DRAS) sends commands over the public Internet to the end-user. The demand response controller on the end-user side actuates over power elements such as HVAC, lighting, pumps, etc. essentially reducing the load in a participating facility.

The successful implementation of autoDR programs will make the power grid much more flexible and efficient at the expense of added system complexity. For instance, Fig. 23 shows a situation where the total electricity generation is insufficient for supplying the load demanded from the grid. This is a plausible scenario as the penetration of renewables increases resulting in an increase in the variability on the supply side. The figure inset exemplifies how automated demand response would actuate on the demand side to compensate for the decrease in the generation. In order to determine how much load needs to be shed, the DRAS requires the forecasting of both generation and the load before the autoDR event. Therefore, in both storage and demand response scenarios, more accurate forecasts are paramount.

Load forecasting is a different subject than solar forecasting but many of the techniques reviewed in this paper can be and are regularly used for the forecasting of power demand. All the stochastic models discussed in Section 3 and the AI techniques discussed in Section 4 are general techniques for the forecasting of time-series and can be readily applied to the forecast of electrical power loads (for more on ADR, please see Ref. [277] and references within).

<table>
<thead>
<tr>
<th>Horizon [min]</th>
<th>RMSE</th>
<th>Forecast skill, s</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Model 1</td>
</tr>
<tr>
<td>30 min</td>
<td>84.70</td>
<td>0.12</td>
</tr>
<tr>
<td>60 min</td>
<td>97.40</td>
<td>0.14</td>
</tr>
<tr>
<td>90 min</td>
<td>109.00</td>
<td>0.23</td>
</tr>
<tr>
<td>120 min</td>
<td>118.00</td>
<td>0.24</td>
</tr>
</tbody>
</table>

11. Conclusion

Solar forecasting methods and applications are covered in this work. From the description of the numerous approaches presented, it is clear that some level of predictive success has been experienced for a number of different spatial and temporal resolutions. Although this research field is incipient, a robust number of techniques and best practices have been developed in the past few years.

Regressive methods, which take advantage of the correlated nature of the irradiance observations tend to work well in both data-poor and data-rich environments. In data-poor environments, one might only have access to historical point sensor or power output data, in which case endogenous stochastic methods can be used such as AR, MA, ARMA and/or ARIMA. On the other hand, in a data rich environment one may have access to a wide number of additional data such as the time evolution of observations of physical quantities. In these environments, exogenous inputs are included through a careful input selection process. These stochastic learning methods have been applied to a wide variety of time horizons ranging from intra-hour to yearly averages as explained in Section 3.

Artificial Neural Network (ANN) modeling offers improved non-linear approximator performance and provides an alternative approach to physical modeling for irradiance data when enough historical data is available. ANNs are generic non-linear approximators that deliver compact solutions for several non-linear, stochastic and multivariate problems. Like regressive methods, ANNs perform well in both data-rich/poor environments and are not typically temporally limited. These techniques have successfully modeled irradiance on intra-hour to yearly time horizons as demonstrated in Section 4.

Satellite imaging provides an alternative to expensive ground based pyranometer networks and allows for the forecasting of irradiance in environments where no other data is available. Elaborate physical RTMs or straightforward regression techniques may be used in conjunction with satellite data to provide irradiance forecasts on time horizons ranging from intra-hour to intra-day (see Section 5). These methods suffer from temporal and spatial limitations due to satellite sampling frequency and limits on spatial resolution of the satellite images.

Numerical Weather Prediction also allows for modeling of the temporal progression of cloud cover in locations were no extensive ground networks are available. NWP models are scalable due to their discrete nature, and can model regional or global atmospheric
References


Boileau E. Guerrier B. Comparison of model statistics saisonniers et non saisonniers en météorologie solaire. La Météorologie 1979;VI:115–30.


Abbreviations

ADALINE: Adaptive Linear Neuron
AI: Artificial Intelligence
AM: Airmass
ANN: Artificial Neural Network
AR: Auto-Regressive
ARIMA: Auto-Regressive Integrated Moving Average
ARIMAX: ARIMA-eXogenous
ARIMA: Auto-Regressive Moving Average
ARMAX: ARIMA-eXogenous
ASCE: American Society of Civil Engineers
autoDR: automated Demand Response
AVHRIR: Advanced Very High Resolution Radiometer
BPL: Back Propagation-Learning
BVE: Barotropic Vorticity Equation
C3S: Clouds and the Earth Radiant Energy System
CFL: Courant, Friedrichs and Lewy
CPV: Concentrated Power
CSF: Concentrated Solar Power
DHI: Diffuse Horizontal Irradiance
DNS: Direct Normal Irradiance
ECMWF: European Centre for Medium-Range Weather Forecasts
ESRA: European Solar Radiation Atlas
ESSA: Environmental Science Services Administration
FPGA: Field Programmable Gate Array
GA: Genetic Algorithm
GFS: Global Forecast System
GHI: Global Horizontal Irradiance
GOES: Geostationary Operational Environmental Satellites
GPS: Global Positioning System
HRRR: High Resolution Rapid Refresh
HS: Hybrid System
IEA: International Energy Agency
ISO: Independent System Operator
KNN: kNearest Neighbor
LARS: Least Absolute Shrinkage and Selection Operator
MA: Moving Average
MDADLINE: Many ADALINES
MAE: Mean Absolute Error
MASE: Mean Absolute Percent Error
MBE: Mean Bias Error
MLP: Multi-Layer Perceptron
MOS: Model Output Statistics
MR: MADALINE adaptation Rule
NAM: North American Mesoscale
NAR: Non-linear AR
NARMAX: Non-linear ARMA-eXogenous
NARX: Non-linear AR-eXogenous
NASA: National Aeronautics and Space Administration
NCAR: National Center for Atmospheric Research
NCEP: National Centers for Environmental Prediction
NOAA: National Oceanic and Atmospheric Administration
NWP: Numerical Weather Prediction
PIV: Particle Image Velocimetry
PLeo: Polar LEO
PV: PhotoVolaic
R: Mean Square Error
RTM: Radiative Transfer Model
RUC: Rapid Update Cycle
SSLEO: Sun-Synchronous LEO
TIRDS: Television InfraRed Observation Satellite
TLU: Threshold Logic Unit
TSI: Total Sky Imager
WRF: Weather Research and Forecasting
XOR: eXclusive-OR
PSSLEO: Polar SSLEO