Solar Energy 168 (2018) 60–101

Contents lists available at ScienceDirect

Solar Energy

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

Review

History and trends in solar irradiance and PV power forecasting: A preliminary assessment and review using text mining

Dazhi Yang[a,*], Jan Kleissl[b], Christian A. Gueymard[c], Hugo T.C. Pedro[b], Carlos F.M. Coimbra[b]

[a] Singapore Institute of Manufacturing Technology, Agency for Science, Technology and Research (A*STAR), Singapore
[b] Center for Renewable Resources and Integration, Center for Energy Research, Department of Mechanical and Aerospace Engineering, University of California, San Diego, CA, USA
[c] Solar Consulting Services, Colebrook, NH, USA

ARTICLE INFO

Keywords:
Text mining
Solar forecasting
Review
Photovoltaics

ABSTRACT

Text mining is an emerging topic that advances the review of academic literature. This paper presents a preliminary study on how to review solar irradiance and photovoltaic (PV) power forecasting (both topics combined as "solar forecasting" for short) using text mining, which serves as the first part of a forthcoming series of text mining applications in solar forecasting. This study contains three main contributions: (1) establishing the technological infrastructure (authors, journals & conferences, publications, and organizations) of solar forecasting via the top 1000 papers returned by a Google Scholar search; (2) consolidating the frequently-used abbreviations in solar forecasting by mining the full texts of 249 ScienceDirect publications; and (3) identifying key innovations in recent advances in solar forecasting (e.g., shadow camera, forecast reconciliation). As most of the steps involved in the above analysis are automated via an application programming interface, the presented method can be transferred to other solar engineering topics, or any other scientific domain, by means of changing the search word. The authors acknowledge that text mining, at its present stage, serves as a complement to, but not a replacement of, conventional review papers.

1. Introduction: Towards a new reviewing paradigm

The history of solar irradiance forecasting can be said to have started in the late 19th- and early 20th-century when numerical weather prediction (NWP) began. It is remarkable how pyrheliometers—the primary instrument to measure direct normal irradiance (DNI), still in common use today as a reference instrument—had already been developed and employed as a forecasting tool by then (Marvin and Kimball, 1926). However, it was not until the advent of mainframe computers and simulations that computation time was reduced to less than the forecast horizon. Today, solar irradiance forecasting and photovoltaic (PV) power forecasting (both referred to as "solar forecasting" in what follows) receive unprecedented attention from various scientific communities. This is because of the importance of forecasting the variability of solar and wind power for their grid integration, which constitutes a major challenge to a successful transformation of the conventional fossil fuel-based energy sector into a 100% renewable one. To give perspective, Google Scholar searches for "solar irradiance forecasting" and "PV power forecasting" return 15,700 and 6340 results for the year 2016 alone.

Considering this abundant literature on solar forecasting, many review papers have been written in recent years. The primary purpose of review papers is to familiarize students and researchers with a relatively new topic and facilitate the use of a number of new and powerful tools. A list of recent review papers on solar forecasting is shown in Table 1. Reviews compile, summarize, critique, and synthesize the available information on a subject (Suter, 2013). Despite the obvious benefits of reviews, they nevertheless have three main drawbacks:

1. The number of references considered in each review is still small relatively to the total available publications on the subject.
2. It is often unclear what methods review authors applied to search the literature, identify publications, extract information, and generate insights (Suter, 2013).
3. Since each review is only read by a handful of scientists (authors, reviewers, and possibly journal editors) before its publication, the content may be biased and/or subjective.

Analogically speaking, review papers behave like local optima in an optimization problem, while actually the global solution is sought. As in optimization, there are ways to escape from the local optima, but it often takes years of experience before a reader can critically interpret

* Corresponding author.
E-mail addresses: yangdz@simtech.a-star.edu.sg, yangdazhi.nus@gmail.com (D. Yang).

https://doi.org/10.1016/j.solener.2017.11.023
Received 23 October 2017; Accepted 9 November 2017
Available online 03 February 2018
0038-092X/

Table 1
Review papers on solar forecasting. The number of citations is taken from Google Scholar at the time of manuscript submission.

Review	Title	Journal	#Refs.	#Pages	#Citations
Barbieri et al. (2017)	Very short-term photovoltaic power forecasting with cloud modeling: A review	Renewable & Sustainable Energy Reviews	94	22	7
van der Meer et al. (2017)	Review on probabilistic forecasting of photovoltaic power production and electricity consumption	Renewable & Sustainable Energy Reviews	140	29	1
Voyant et al. (2017c)	Machine learning methods for solar radiation forecasting: A review	Renewable Energy	105	14	18
André et al. (2016)	Predictive spatio-temporal model for spatially sparse global solar radiation data	Energy	29	10	2
Antonanzas et al. (2016)	Review of photovoltaic power forecasting	Solar Energy	151	34	44
Raza et al. (2016)	On recent advances in PV output power forecast	Solar Energy	123	20	28
Kashyap et al. (2015)	Solar radiation forecasting with multiple parameters neural networks	Renewable & Sustainable Energy Reviews	65	11	19
Qazi et al. (2015)	The artificial neural network for solar radiation prediction and designing solar systems: A systematic literature review	Journal of Cleaner Production	54	12	38
Ren et al. (2015)	Ensemble methods for wind and solar power forecasting–A state-of-the-art review	Renewable & Sustainable Energy Reviews	55	10	35
Wan et al. (2015)	Photovoltaic and solar power forecasting for smart grid energy management	CSEE Journal of Power and Energy Systems	82	9	54
Law et al. (2014)	Direct normal irradiance forecasting and its application to concentrated solar thermal output forecasting—A review	Solar Energy	165	21	46
Diagne et al. (2013)	Review of solar irradiance forecasting methods and a proposition for small-scale insular grids	Renewable & Sustainable Energy Reviews	50	12	170
Inman et al. (2013)	Solar forecasting methods for renewable energy integration	Progress in Energy and Combustion Science	293	42	265
Kleissl (2013)	Solar energy forecasting and resource assessment	Book	–	–	124

and synthesize these reviews to derive an objective assessment of the state-of-the-art.

In this paper, an assistive method—text mining—is primarily considered as a potential replacement for, or addition to, conventional literature reviews. Since text mining is an automated process of deriving information from text, it is not limited by the amount of input data, thus providing a remedy for the first aforementioned drawback. In each of the sections below, the methods used to collect, group and analyze publications are elaborated with justification. Such elaboration is believed to improve the transparency of the present results. In turn, this should provide greater assurance of the quality of the review process, and hence close the second gap mentioned earlier. Lastly, to reduce the unavoidable biases in any review, a group of domain experts—five associate/subject editors of Solar Energy[1] on the subject of solar resources & energy meteorology—are interpreting the text mining results and co-writing this paper. Furthermore, Google Scholar search results are herein considered. Since Google Scholar ranks a publication based on (i) where it was published, (ii) who it was written by, as well as (iii) the count and recency of its citations,[2] the search results essentially reflect the prevailing confidence in popularity and publication quality (as suggested by crowdsourcing). Based on this assessment, the combination of Google Scholar data and supervision from domain experts is expected to mitigate the third drawback.

2. Introducing a new toolkit for literature review

2.1. Working with Google Scholar data

Google Scholar is one of the most important free academic search engines (Ortega and Aguillo, 2014), and often provides a more comprehensive coverage of resources in various scientific disciplines as compared to Web of Science or Scopus (Harzing, 2013). By mining and analyzing the environment of a large number of publications (e.g., titles, authors, abstracts, citations, and Google Scholar profiles), valuable information and insights on an academic field can be obtained.

Much research has been done in various fields using Google Scholar data. For instance, Chen et al. (2017) collected more than 400,000 Google Scholar profiles across various disciplines and analyzed the demography of these scholars. A co-authorship network was built to study the collaboration among authors and its resulting link to citation metrics. It was found that the ranking of a page is strongly correlated with the h-index.[3] From a different perspective, Shariff et al. (2013) utilized Google Scholar to help physicians to retrieve clinical evidence and to guide the care of their patients. In the field of knowledge management, Google Scholar was used to discover growing, stable and declining research trends (Serenko and Dumay, 2015). Google Scholar data has also been used in solar engineering. Yang (2016) compared citations of 15 papers on irradiance transposition modeling through years, and filtered out the less-cited models for that study.

Despite all its potentials and benefits, Google Scholar has its downside: the lack of transparency is the main reservation of bibliometricians to use it as a research evaluation database (López-Cózar et al., 2014). Because Google Scholar automatically retrieves, indexes, and stores any form of text-based scientific material (paper, presentation slides, or even personal memo) uploaded by an author without much quality control, information such as citation counts may be inflated. López-Cózar et al. (2014) performed an experiment by uploading several false papers with abundant citations of publications from their

[1] For clarity, journal and author names (only when not in a citation) are noted with Small Caps.

[2] https://scholar.google.com/intl/en/scholar/about.html.

[3] The h-index was proposed by Hirsch (2005) to characterize the scientific output of a researcher. It is defined as the number of papers with citation number ⩾h. Google Scholar separately calculates the h-index of all scientists based on their whole career and on the latest 5-year period.

lab. An outburst in the number of citations on those scholars' profiles was then observed. Based on this evidence, one needs to be cautious when using Google Scholar data. It appears most appropriate to combine Google Scholar with other sources of data whenever possible.

In this contribution, Google Scholar data is primarily used for knowledge discovery on the technology infrastructure—a term used by Delen and Crossland (2008)—of solar forecasting, which includes information on journals & conferences, authors, publications and organizations. Understanding the technology infrastructure of an area facilitates several things, including but not limited to:

1. Decision making for manuscript submission, by identifying appropriate journals and conferences with high impacts for a research topic (e.g., International Journal of Forecasting is a great journal, however, it does not currently publish many solar forecasting papers, for unknown reasons).
2. Research collaborations, by knowing top researchers who have similar research interests as the author (e.g., someone who has never heard of Richard Perez or Elke Lorenz is most likely new to solar forecasting).
3. Building up a reading list, by selecting well-cited papers (scientists cannot read all papers, but the internet can filter the good ones for them).
4. Career development, by becoming aware of organizations and institutes that have programs related to one's own research topic, hence potentially benefiting one's future job search.

The technology infrastructure of solar forecasting will be discussed in detail in a later section.

2.2. Working with full texts

Even though Google Scholar contains abstracts of papers, which by definition summarize the main content of the papers, it is usually not sufficient for an in-depth understanding of a subject. Analyzing full texts is therefore essential to perform a comprehensive review. A direct way to prepare full texts for text analytics consists in parsing PDF (portable document format) files. However, converting unstructured text data—if it is text data at all, since many PDFs (particularly for old documents) are in fact scanned images—to structured data is not an easy task. Similarly, any text included in figures or pictures cannot be extracted. Although tools such as **Xpdf** and **Poppler**[4] can read a majority of recent PDF files, problems such as translating some encoding marks, removing headers and footers, translating ligatures, or distinguishing end-of-line breaks from hyphens are still difficult to handle. The "perfect" tool that would be able to easily and transparently perform all this procedure does not seem to exist yet. Instead, parsing PDFs of scientific publications usually requires trial-and-error, check-and-modify, and case-by-case programming. These intermediate steps are often time consuming and cannot be circumvented. However, although the parsed PDFs are usually not perfect, they still carry a lot more information than the papers' abstracts.

Fortunately, many scientific publishers recognize the importance of text mining and provide an application programming interface (API) for researchers to access different levels of information. While most publishers only provide APIs on metadata for their documents, Elsevier developed a platform for its authenticated developers to access full text content.[5] This policy aligns well with the present text mining needs. In this study, the full texts of 249 journal papers are accessed via the Elsevier API and used to review the topic of solar forecasting. As opposed to parsing PDF—approach that was used initially for this work—the full text content obtained via API is much more amenable to work with.

2.3. Text mining's role in reviews

Text mining is a rapidly developing field, at least comparable in importance to solar engineering.[6] Text mining combines techniques in data mining, machine learning, natural language processing, information retrieval, and knowledge management, to solve the crisis of information overload in today's world (Feldman and Sanger, 2007). Although the application under scrutiny, i.e., reviewing the academic literature, is merely a small area in text mining, it still opens up a new paradigm and is expected to bring major advances to literature review methods in the very near future. Further exploiting text mining techniques has the potential to:

1. Construct a relatively unbiased technology infrastructure.
2. Generate centralized, domain-specific dictionaries to promote the universal acceptance of terminologies.
3. Discover new concepts, approaches, and methods.
4. Cluster and classify the main themes in a specific academic domain.
5. Associate technology infrastructure, themes, and methods.
6. Summarize research directions and topics chronologically.
7. Project future research directions and topics.

Since this paper is the first of this kind in solar engineering, the focus is placed here on the first three tasks. A brief introduction to text mining is provided in Section 3. The technology infrastructure is developed in Section 4, based on the top 1000 results returned by a Google Scholar search. All four points enumerated in Section 2.1 are discussed there. Section 5 generates a list of frequently-used abbreviations in solar forecasting by mining the full texts of 249 papers obtained from ScienceDirect. Frequently-used abbreviations are annotated with technical details and interpreted, based on a classification of concepts of solar forecasting. Section 6 is concerned with a number of emerging technologies. Six publications are handpicked because of their perceived strong potential impact in the future. Further analysis, consisting of keyword analysis and topic modeling, is then performed on them. The concepts, approaches, and methods of these emerging technologies are thus studied in depth.

3. Introduction to text mining sequence

In the recent book by Kwartler (2017), a text mining task is decomposed into six steps:

1. Define the problem and goal.
2. Identify and collect the text.
3. Organize the text.
4. Extract features.
5. Analyze.
6. Reach an insight or recommendation.

The concept and importance of text mining—in the context of creating a review on an academic topic—are introduced in what follows by describing these six steps.

3.1. Problem definition and goal setting

Like with any other data analytics task, it is necessary to have a problem before one can have a solution. Since the goal of a review paper is to help the readers acquire the essential matter on a specific

[4] Software packages and functions are noted using **bold font** in this paper.

[5] https://www.elsevier.com/about/our-business/policies/text-and-data-mining.

[6] Google Scholar searches returns 2.7 million and 2.5 million results for "text mining" and "solar engineering", respectively. In contrast, regular Google searches return 12 million results on text mining and 3 million on solar engineering.

subject, the problem definition (elaborated below) should evolve around this goal.

The first and foremost question that a novice ought to ask is: where to start? A parallel can be drawn with how to deal investments in the financial world: it is not prudent to start without a thorough knowledge of all intricacies, resulting from substantial market research. Before diving down into surveys of the state-of-the-art, the specific academic environment and related paradigm must be understood first. Authors often get rejections from a journal due to misalignment with that journal's scope. Under other circumstances, it is also important to know who the domain experts are, and which papers are already popular in that field, so that a benchmark for a planned piece of research can be set. To that end, the first general question that must be formulated here is: (Q1) what are the relevant journals & conferences, who are the leading researchers, which publications are influential in solar forecasting, and which organizations are actively pursuing solar forecasting research?

Once the answers to the above question are known, a review author may want to know the important concepts involved in solar forecasting. The smallest entity to illustrate a concept is a single word. For example, when one is interested in forecasting methods, the word "persistence" refers to a forecasting method that assumes the forecast is the same as the previous observation (or an observation during a similar time of the recent past). Finding out important words can be done by counting the appearances of specific words. It is logical to assume that an important word would appear more often than a less important word. However, more than often, a word is not sufficient to describe a concept, so a phrase is required at least. When an important phrase needs to be mentioned multiple times, authors often use abbreviations. For instance, "ANN" can replace the phrase "artificial neural network" if needed. Hence, a logical second question is: (Q2) which are the most important and frequent abbreviations in solar forecasting? By answering this question, a knowledge web can be formed. With a little reading, various concepts discovered through abbreviations can be classified into groups.

Words and abbreviations describe concepts, but further elaboration is required for a more in-depth understanding. Most concepts used in solar forecasting are well exploited, and a short description of them may be found on Wikipedia[7] or other online references, but advanced or emerging concepts may require more attention due to lack of a good summary. Undoubtedly, the best way to understand an emerging concept is to read the original publication in detail. However, at the exploratory stage of a research project, detailed reading may be time consuming and not efficient. After emerging concepts have been found, a third question is therefore: (Q3) what are the keywords, phrases, or topics associated with those concepts?

To investigate the above three questions with reliable data, three different datasets have been assembled here: 1000 abstracts from Google Scholar search results, 249 full texts from ScienceDirect, and six recent articles (published in 2016 or 2017).

3.2. Data choice and collection

To collect data for text mining, the proper selection of text is important. The simplest way to collect text data for reviews is by selecting PDF files that are related to the problem definition, but this process requires manual download. When an API is available, it is straightforward to collect text via the API. Other times, some form of web scraping is needed (see reviews by Singh and Vikas, 2014; Kausar et al., 2013). Web scraping often requires customization of scripts due to possible HTML variations. In this work, all three types of channels (PDFs, APIs, and web scraping) are used to collect data. Hand-scraped HTML files are used for Q1, full texts obtained via API for Q2, and carefully selected PDF files for Q3. Although it is also possible to use APIs for Q3, the present study considers working with PDF files, and provides some discussion on the difficulties one faces when working with PDF files.

3.2.1. Collection of Google Scholar data

Although Google Scholar has an extensive coverage of literature, it does not provide any API, probably due to the expected overwhelming requests and potential abuse of the data. Furthermore, only the first 1000 results are available for viewing. Therefore, some manual work is expected when working with Google Scholar data. Downloading HTML files for the first 1000 results that were returned by the search term "solar + irradiance + PV + power + forecasting" in this paper, took about 5 min. One can also repeat the procedure with more detailed search words once the general technological infrastructure has been established and a more specific sub-domain needs to be expanded. Considering that the solar forecasting literature is fast expanding and the search results may vary based on geographical location, the data presented here is only representative of the search made from Singapore on 2017-07-23.

Before the HTML files can be used, their content need to be processed, since the search results from Google Scholar are often incomplete. The word *incomplete* may refer to two different concepts in the present context:

1. Texts stored in HTML files may be incomplete due to browser display constraints. For instance, the journal named Progress in Photovoltaics: Research and Applications often also appears as "Progress in ⋯"; author names and abstracts may be omitted after a certain length and replaced with "⋯", etc.
2. Incomplete metadata (e.g., author-identified keywords are not shown in Google Scholar).

To that end, digital libraries that host the respective publications are used to retrieve the metadata. In this paper, eight digital libraries, namely, ScienceDirect (containing 483 out of the retrieved 1000 publications), IEEE Xplore Digital Library (276/1000), Wiley Online Library (27/1000), Institute of Engineering and Technology (20/1000), Multidisciplinary Digital Publishing Institute (17/1000), Springer (16/1000), Hindawi (15/1000) and Taylor & Francis Online (4/1000), are considered here. They jointly cover more than 85% of the first 1000 search results. After parsing the HTML files, fields such as title, author, journal, year, URL or citations are consolidated into a data table. Subsequently, the metadata pulled using the respective APIs from the digital libraries are used to modify the data table by replacing or adding fields.

3.2.2. Collection of ScienceDirect full text data

Google Scholar excels at giving broadness and objectiveness to a search. It however lacks the ability to provide in-depth content, such as the text in PDF files. The full texts can be accessed from the respective digital libraries, such as ScienceDirect or IEEE Digital Library. As mentioned in Section 4 (see below), ScienceDirect is the most "popular" library for solar forecasting. Hence, it is interesting to conduct an advanced search on ScienceDirect, namely, "TITLE (forecast AND NOT wind) and (solar irradiance OR PV power) [All Sources (Computer Science, Energy, Environmental Science, Mathematics, Physics and Astronomy)]". A total of 307 results were returned. All publications without any author, e.g., editorials, newsletter or communications, were eliminated. Similarly, some journals like Solar Energy Materials and Solar Cells or Journal of Atmospheric and Solar-Terrestrial Physics contained the search words but were found irrelevant and discarded. After these filtering steps, a total of 249 full texts remained. They were then analyzed to generate a list of frequently-used abbreviations for solar forecasting, as reported in what follows.

[7] Wikipedia is not 100% reliable, but it can be very useful.

3.2.3. Collection of full texts on emerging technologies

In the present case, the concept of emerging technologies refers to a selection of articles published in 2016 or 2017. As these publications may have yet to receive the attention they deserve, a selection based only on Google Scholar page ranks and citations is not suitable. Consequently, the best alternative is thought to be voting. In the present case, a list of 33 recent solar forecasting publications from Solar Energy[8] was first built by the lead author. Subsequently, each author independently nominated 10 top papers. The best paper was given a score of 10; the tenth paper was given a score of 1; and the unselected papers were given a score of 0. Once the ranking data is collected, a linear ranking method (Alvo and Yu, 2014) is used to consolidate the results. The linear ranking method considers the *mean rank* $\boldsymbol{m} = (m_1,\ldots,m_{33})^{\mathrm{T}}$ of the publications. For the ith publication

$$m_i = \sum_{j=1}^{33!} n_j \nu_j(i)/n, \tag{1}$$

where ν_j, $j = 1,2,\ldots,33!$ represents all possible rankings of these 33 publications; n_j is the frequency of occurrence of ranking j; $n = \sum_{j=1}^{33} n_j$; and $\nu_j(i)$ is the score given to publication i in ranking j. Finally, six publications with highest mean ranks are selected as the text data for analyses on emerging technologies. An example of application of this process is described in Section 6.

3.3. Organizing the text—bag-of-words

There are two types of text mining: (1) bag-of-words (BoW) and (2) syntactic parsing (Kwartler, 2017). BoW is not concerned with word order or grammatical word type. Hence, one of its advantages is that it is not computationally expensive. BoW can be represented with a document term matrix (DTM), where each row represents a document and each column represents a word or phrase. The matrix representation of BoW aligns nicely with a machine learning framework: instead of dealing with semantics, it translates words into numbers.

As an alternative to BoW, syntactic parsing performs text mining based on syntax. Syntactic parsing respects the various parts of speech. Therefore, it can identify some grammatical aspects of the words, such as nouns, verbs and adjectives. It is obvious that syntactic parsing captures more information than the BoW methodology. Nevertheless, the simpler BoW approach will be the focus of this work.

3.4. Feature extraction

Creating features means that text needs to be preprocessed before any specific analytical step. In other words, before the desired BoW can be formed, preprocessing is usually needed, since the input texts can be difficult to manipulate in their untidy raw format. Some commonly-used preprocessing steps include tokenization, stop-word removal, whitespace removal, punctuation removal, upper-to-lower case conversion, stemming, regular expression-based filtering, and, sometimes, synonyms conversion. Most of the above-mentioned preprocessing steps are self-explanatory, but details are provided here on tokenization, stop-word removal, stemming, and regular expression-based filtering for the reader's benefit.

Tokenization breaks text into tokens. Tokens can be words (also known as unigrams), phrases (bigrams, trigrams, …, n-grams), or even sentences. *Stopwords* refer to the most common words in a language, which usually do not contribute to text mining. Examples of stopwords include "we", "and", "of", "that", etc. The stopword list can be built based on a particular text mining task. For example, if the word "forecast" systematically appears in all documents, it can be set as a stopword and removed during preprocessing. *Stemming* reduces inflected words to their word stem. For example, words "forecasts" and "forecasting" have the same word stem, "forecast". A *regular expression* (Thompson, 1968) is a pattern that describes a set of strings. It searches specific strings embedded in text, so that operations such as replacement and filtering can then be applied. It is a very powerful and most useful tool for text mining. However, dealing with regular expressions involves a rather steep learning curve. The reader is referred to the book by Krause (2017) and other online sources[9] for tutorials on regular expression. Most of the commonly-used preprocessing tools are part of the **tm** package in **R** (see Feinerer and Hornik (2017) and Feinerer et al. (2008), for a list of functions). As the goal of preprocessing is to tidy the input texts, its sequence may vary with raw text and application.

HTML files require additional steps even before preprocessing. The HTML document structure needs to be known, which implies determining what the available node names are and what they contain. Because HTML files can be quite long, scanning them line-by-line may not be practical. The **R** package called **rvest** is particularly useful in this situation. This package is a wrapper around several related **R** packages, and makes it easy to download and manipulate HTML. In particular, the function **html_nodes** is very useful to quickly extract pieces out of HTML documents using **XPath** and **css** selectors.

One last thing to discuss here is *ligature*, which is a particular problem when working with PDFs. In typography, a ligature occurs where two or more graphemes are joined as a single glyph. This happens, for example, when letters "f" and "l" appear together, as in "fl". Ignoring ligature translation affects the accuracy of any subsequent text analysis. Consider the task of matching the abbreviation "VOF" to its long form. This is not possible without ligature translation, even when the term "variational optical flow" is correct, since the letter "f" is not recognizable. In this work, the **R** package called **tau** is used to translate all ligatures to alphabets.

3.5. Analyzing the extracted features

After the text is preprocessed and features are extracted, analytical methodologies need to be applied to gain insights, recommendations, or to confirm existing knowledge about the problem (Kwartler, 2017). The analysis in a text mining context can be rather simple, e.g., search for the number of occurrences of the word "forecast", or involve sophisticated algorithms, such as *unsupervised ontology induction*. Whereas there are many analytical methods for text data, only three methods are investigated here because they are thought most relevant to the present application, namely, reviewing an academic topic:

1. Analyzing word frequency.
2. Analyzing relationships between words.
3. Topic modeling.

All three methods listed above revolve around a central question in text mining, namely, how to quantify what a document is all about. The method for analyzing word frequency is built upon the belief that important concepts appear more often. By counting the number of appearances of each word in a document—this is known as the term frequency—words with highest frequencies can be selected and interpreted. However, in our present context, one can expect that words such as "solar" and "forecasting" would appear many times in all documents. Listing these common words may not be most meaningful. To resolve this issue, one can either add these common words to the stopword list, or use the concept of term frequency–inverse document frequency, which measures how important a word is to a document with respect to other documents in the corpus.

[8] Undoubtedly, there are important contributions to solar forecasting published in other journals. Solar Energy was chosen because it is the most popular journal for solar forecasting as will be shown in Fig. 1 later.

[9] In particular: http://www.rexegg.com/ and https://stat.ethz.ch/R-manual/R-devel/library/base/html/regex.html.

In terms of analyzing relationship between words, n-grams and correlations are widely used. An n-gram is a sequence of words that occur consecutively in text. In the two-word case, the sequence of words is known as a bigram; in the three-word case, the sequence is known as a trigram. Counting n-grams provides an understanding on how often a word X follows a word Y. Similar to counting n-grams, correlation is another measure of co-occurrence of words in near proximity, e.g., in the same sentence or paragraph.

Lastly, topic modeling is a method for unsupervised classification of documents. Given a collection of documents, it is of interest to observe natural groups in them. For example, a particular topic on solar forecasting is cloud modeling. It would be useful to check the relevance of a solar forecasting publication to this topic. Latent Dirichlet Allocation (LDA) is a popular choice of topic modeling. Under the LDA framework, each document is treated as a mixture of topics, and each topic consists of a mixture of words.

3.6. Interpreting the text mining results

Whether it is a result of simple counting or some complex analysis, the result needs to be directly related to the goals set earlier. Whereas the results related to Q1 are rather straightforward to interpret, good interpretations on results related to Q2 and Q3 require domain knowledge. Therefore, throughout the remaining part of this paper, scattered ideas about solar forecasting are consolidated based on the text mining results. Even though it is almost surely impossible to discuss all concepts in one document, this paper is believed to be comprehensive enough to motivate research on all major aspects in the area of solar forecasting.

4. Solar forecasting technology infrastructure

In this section, the HTML files downloaded from Google Scholar, as described in Section 3.2, are used to discover the technology infrastructure of solar forecasting.

4.1. Journal & conference infrastructure

With the complete data table, an immediate interest would be to identify the top journals & conferences through simple counting. Fig. 1 reveals the top 20 journals & conferences using a ranking based on the number of appearances in the first 1000 results returned by Google Scholar. Solar Energy ranked first, with a total of 121 papers listed, followed by other journals, such as Renewable Energy (67), Renewable and Sustainable Energy Reviews (66), Energy Conversion and Management (43), Energy (43) or Applied Energy (41). Remarkably, all top six journals are Elsevier journals. It should be noted that this ranking does not reflect the quality of the journals, nor correlates with their actual impact. Journal impact factors are indicated in Fig. 1 using colors. Instead, this ranking reflects the journals' interest in publishing solar forecasting papers and the researchers' interest in submitting their work to these journals.

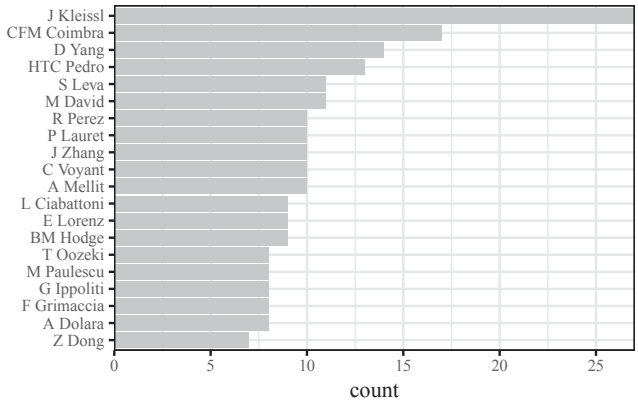

Fig. 2. Top 20 solar forecasting researchers ranked based on number of appearances in the first 1000 results returned by Google Scholar.

4.2. Author infrastructure

A similar counting is performed on the authors as well, as shown in Fig. 2. Based on the particular set of Google Scholar search results described earlier, Jan Kleissl was found to have most appearances, with 27 publications, followed by Carlos Coimbra (17), Dazhi Yang (14) and Hugo Pedro (13). This list may be limited by the search criteria selected here: the search period only covers the recent five years. Indeed, renowned solar forecasters such as Richard Perez or Elke Lorenz had many papers prior to 2012. Nevertheless, this result reflects the current author infrastructure and is useful in identifying active researchers. However, it should again be noted that this ranking does not represent a researcher's achievement. Rather, the ranking only suggests that those researchers who appear in the list may have had more publications in solar forecasting than others during the past five years. A more thorough way to build an author infrastructure would be to consider factors such as citations, h-index, years in the field, among other criteria. The reader is referred to Acuna et al. (2012)—a high-impact Nature paper—for an interesting way of measuring and predicting scientific success.

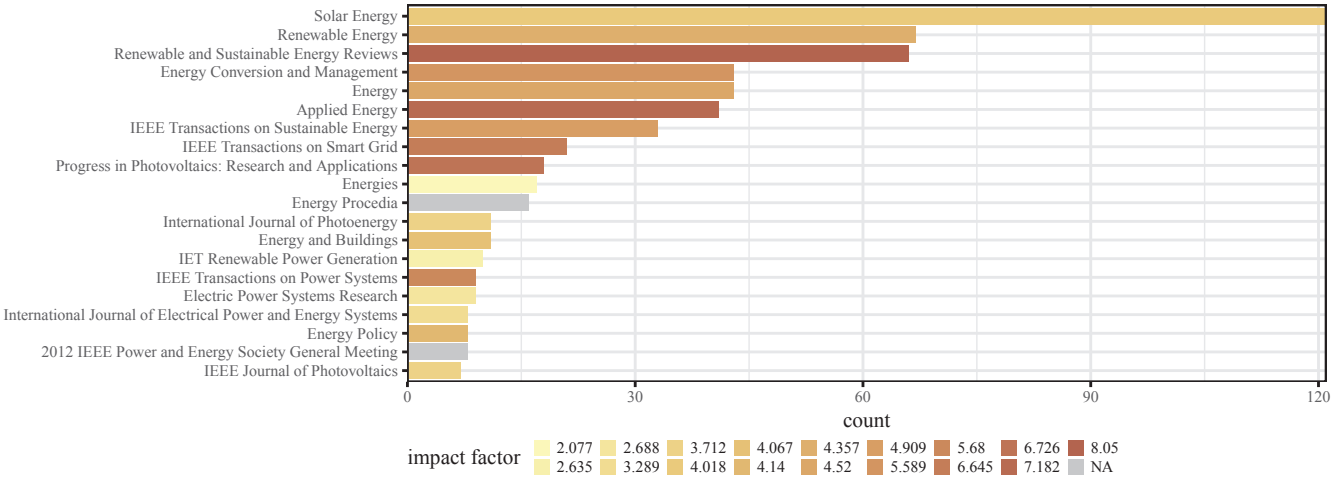

Fig. 1. Top 20 solar forecasting journals & conferences ranked based on number of appearances in the first 1000 results returned by Google Scholar.

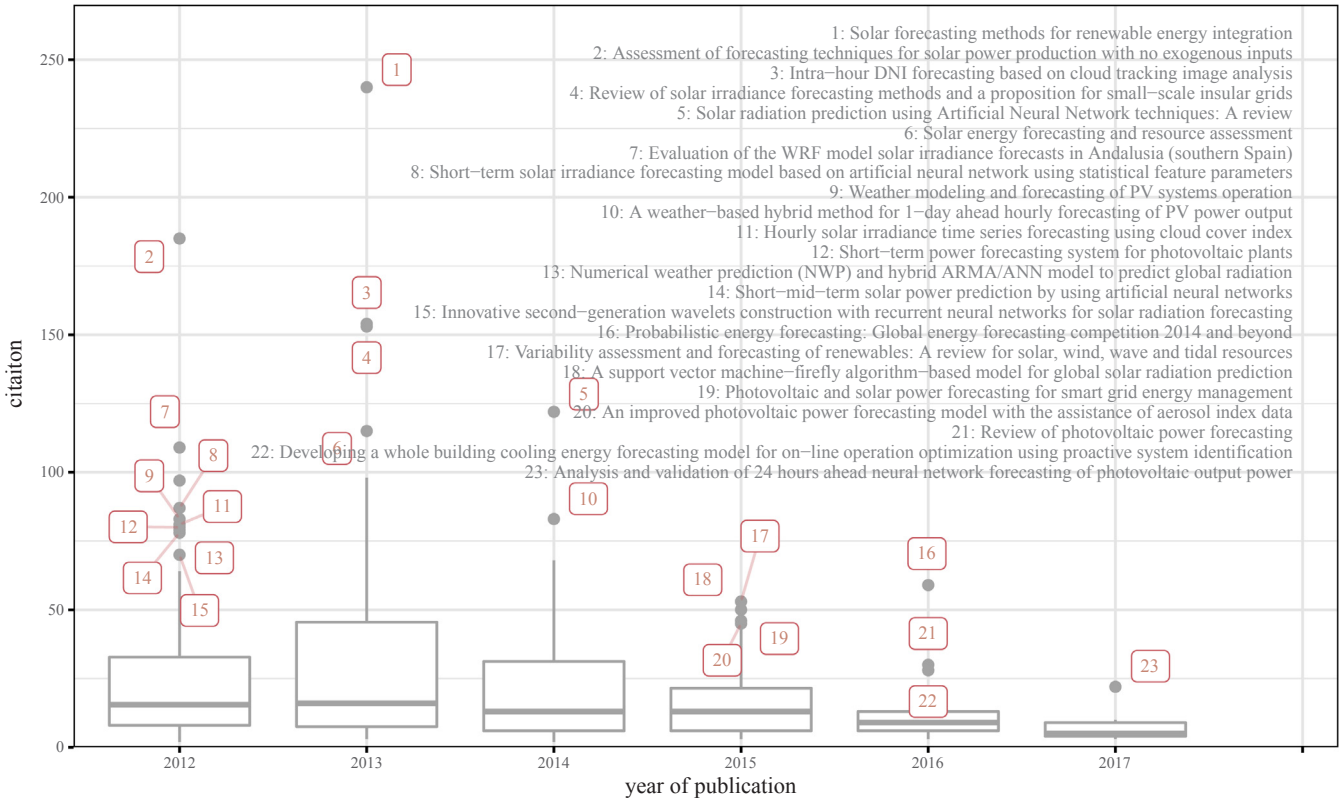

Fig. 3. Box plots of citation numbers for papers published in each year. Titles for the outlier papers (more cited) are printed. See text for details.

4.3. Publication infrastructure

Next, as mentioned in Section 2.1, it is of interest to construct a reading list. Such a list can help someone unfamiliar with the topic to rapidly gain some understanding of the issues and major contributions. Since an article published earlier may receive more citations than a recent publication, the top papers are identified by plotting the number of paper citations against the year of publication. The box plots of citations grouped by year this way are shown in Fig. 3. A standard outlier identification rule is used to detect the more cited papers in each year. More specifically, if the citation number of a paper is larger than the third quartile by at least 1.5 times the interquartile range, it is identified as an outlier. Amongst various top papers, the review paper by Inman et al. (2013) received most citations (242) as of July 2017. Therefore, for any scientist willing to do research on solar forecasting, reading Inman et al. (2013) and other publications shown in Fig. 3 would be a good starting point.

4.4. Organization infrastructure

The last major component of technological infrastructure is its organizations. As Google Scholar results do not contain author affiliation information, the author affiliation needs to be retrieved from the proper online libraries (as discussed at the beginning of this section). By accessing various APIs provided by these online libraries, the author affiliation information is associated with each paper. Fig. 4 shows two groupings of affiliations: (a) by country and (b) by organization. In Fig. 4(a), it is observed that solar forecasting has received world-wide attention during the past five years, with the United States being the biggest technology center in terms of number of publications. In Europe, the top five players in solar forecasting research are Italy, Spain, France, Germany and United Kingdom. In Fig. 4(b), author affiliations are grouped according to the authors' institutions or organizations. Organization names with more than five appearances are printed in the plot. This list reflects organizations that are active in doing solar forecasting research, which in turn provides a good reference for any job seeker or collaboration initiation.

5. Abbreviations for solar forecasting and interpretations

By examining publication titles in Fig. 3, it is found that abbreviations (or acronyms if they are formed by initial letters) are often used in article titles. However, besides "NWP", which is explained to be numerical weather prediction, other abbreviations including "ANN", "ARMA" or "WRF" are not explained, causing confusion to someone who is not familiar with them. The growing use of abbreviations presents challenges not only for human readers but also for computer programs during text mining. Furthermore, as the literature grows, uncontrolled and non-standardized use of abbreviations is often found. Therefore, having a lexicon of commonly used abbreviations for a particular scientific domain is important.

Abbreviations can be thought of as seeds for literature review because abbreviations are motivated by terminology that is frequently used and that consists of long technical terms that are cumbersome to spell out. By examining a list of abbreviations, various concepts of a specific scientific domain can be classified in a relatively quick way, which is otherwise time consuming and often incomplete. Abbreviations retrieved from solar forecasting texts can be used to construct a *dendrogram*, as will be demonstrated further below in this section. A dendogram is a tree structure that shows taxonomic relationships and classifies various important concepts.

Abbreviations can be retrieved from abstracts of articles, but a full-text search is expected to yield more representative results. For the present preliminary work, 249 solar forecasting full texts hosted on ScienceDirect (as described in Section 3.2.2) are considered and analyzed. A text mining algorithm similar to the one proposed by Schwartz and Hearst (2003) is used to retrieve all abbreviations.

Abbreviation handling is one of the challenges in text mining,

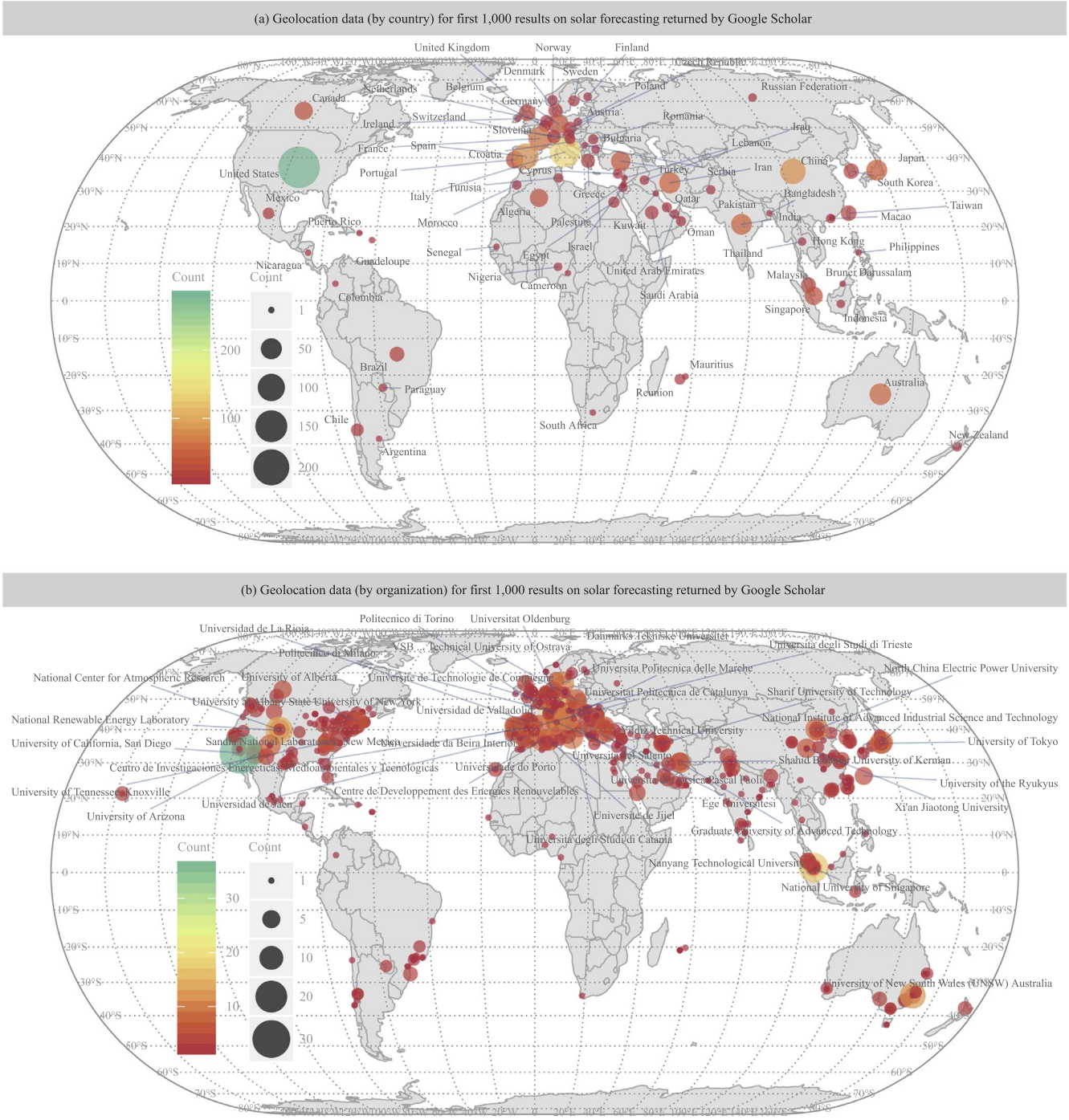

Fig. 4. Geolocations of author affiliations in the first 1000 results returned by Google Scholar (a) grouped by country and (b) grouped by organization.

especially during the construction of lexical ontologies. Although most abbreviations are simply made of word initials, there are many cases for which the matching requires word skip (e.g., "**r**andomized **t**raining and validation set **m**ethod" is abbreviated as "RTM" by Chu et al., 2013), or using internal letters (e.g., "**e**xponen**t**ial **s**moothing" is abbreviated as "ETS" by Hyndman et al., 2008). Training- and learning-based abbreviation identification algorithms often fail, due to the ever-expanding novel use of abbreviations. Consider the abbreviation "RTM": most training-based algorithms would suggest "**r**adiative-**t**ransfer **m**odel", which is a more frequently seen usage in solar engineering. To that end, the tool used here is a neighborhood-search method, which examines the text around an abbreviation for potential match.

The algorithm considers a ⟨short form, long form⟩ pair. Such pairs can be detected by one of two methods: (1) if a nomenclature section is present, the pairs discovered in it are used as is; or (2) the pairs are detected through parenthesis searching. In academic papers, it is common to introduce an abbreviation at its first appearance using parenthesis in a form that is either "short form (long form)" or "long form (short form)". In this specific work about solar forecasting, the latter case is found much more frequently than the former case. Hence, only the latter case is considered in what follows. Regular expressions[10] are used to locate parentheses and to label them as short forms. The corresponding long form is then found by selecting all preceding terms

[10] https://stat.ethz.ch/R-manual/R-devel/library/base/html/regex.html.

```
...solar forecasting applications.  The USI captures images using an upward-facing charge-coupled
device (CCD) image sensor sensing RGB channels...
```

Text sample. from Yang et al. (2014d).

in the same sentence, up to the first non-hyphen punctuation.

In Text sample 1, the short form and long form are "CCD" and "The USI captures images using an upward-facing charge-coupled device", respectively. The matching starts by scanning the long form in reverse order, until the first appearance of the last letter in the short form is found. In Text sample 1, the letter "D" first appears in "device" (case insensitive). The scan continues from the previous cursor position, and loops until all the letters in the short form are found in the long form. The algorithm only has one rule: the match of the character at the beginning of the short form must match a character in the initial position of a word in the long form. If any letter is not found when the long form is completely scanned, the abbreviation does not have a match, and a null is registered.

This algorithm is found to be very effective for the texts analyzed here. However, it has several limitations that may affect its accuracy. Suppose the word "coupled", in the above text sample, is misspelled as "coucpled", with an additional "c" in the middle. The algorithm would then identify "**c**ou**c**pled **d**evice" as the long form, instead of the correct "**c**harged **c**oupled **d**evice". Moreover, it cannot identify the correct short form if the parentheses that contain the short form also contain a citation, e.g., "(CCD, Yang et al., 2014b)". For these reasons, and other potential pitfalls of this automated abbreviation detection algorithm, it is thought that a more focused future study would be beneficial.

Through the above exercise, a total of 1145 unique short forms and 1521 unique long forms are detected. This discrepancy confirms an earlier statement: the uncontrolled and non-standardized use of abbreviation is common in the academic literature. The multiple long forms that match a single short form can be caused either by confusion or by incorrect usage. The list below summarizes some common types of confusion (noted by letter C) and incorrect usage (noted by letter I) seen in the retrieved list of abbreviations:

1. (C) Multiple correct long forms match a single short form, e.g., both "realtime market" and "randomized training and validation set method" match "RTM".
2. (C) Words such as "technique", "model", "method" and "component" confuse abbreviations, e.g., "moving average component (MA)" is abbreviated the same way as "moving average".
3. (C) Missing words from the long form, e.g., "Kolmogorov–Smirnov Integral (KSI)" means "Kolmogorov–Smirnov test Integral", but misses a word.
4. (C) Confusion caused by British and American spelling differences, e.g., neighbour vs. neighbor.
5. (I) Words that are not correctly written, e.g., "autoregressive" is often written as "auto-regressive" or "auto regressive".
6. (I) Wrong words with identical initials as the correct words are used, e.g., "mean average percentage error" instead of the correct "mean absolute percentage error".
7. (I) Wrong words with same stems as the correct words are used, e.g., "least square" instead of the correct "least squares".
8. (I) Creating abbreviations when a well-established abbreviation is available, e.g., "exponential smoothing state space" is abbreviation as "ESSS", whereas "ETS"[11] is the well-accepted abbreviation.

Even though some of the confusion and incorrect usage is obvious, other causes of discrepancy are debatable. To resolve such conflicts, a counting is performed here, based on the frequency of use for each short form. Only the most commonly-used abbreviations are registered. In contrast, those abbreviations with only a single appearance are filtered out. After these two steps, 372 ⟨short form, long form⟩ pairs remain. Lastly, the abbreviations that have little to do with solar forecasting and/or carry little importance, e.g., "alternating current (AC)" or "day ahead (DA)", are removed manually. The final list of frequently-used abbreviations is classified and plotted in Fig. 5 and interpreted in the remainder of this section.

5.1. Error evaluation

By examining the list of frequently-used abbreviations, it is found that forecast evaluations can be categorized into two types: (1) evaluation metrics for point forecast; and (2) evaluation metrics for probabilistic forecast. These metrics are compiled in Tables 2 and 3, respectively.

5.1.1. Error metrics for point forecast

It is apparent from Fig. 5 that the root mean square error (RMSE) is the most widely-used metric in point-forecast evaluation. This popularity may be due to the fact that large errors are particularly undesirable in solar forecasting, so that RMSE, which penalizes large errors, is more appropriate than e.g. mean absolute error (MAE) (Yang et al., 2015d). Nevertheless, MAE and mean bias error (MBE) are very frequently used as well.

Metrics such as RMSE, MAE and MBE are known as scale-dependent errors (Hyndman et al., 2008). As such, they have limitations when comparing forecast accuracies across data with different scales. Therefore, scale-independent error indicators may be more useful in this case. Usually, scale-independent error indicators are represented in percent of an average value, e.g., normalized MAE (nMAE), normalized MBE (nMBE) and normalized RMSE (nRMSE). Alternatively, many authors report percent values for RMSE, MAE and MBE with no change in acronym (e.g., Gueymard, 2014). It is important to note that, although it is common to normalize these error indicators by dividing them with the mean of measurements, many publications in the literature do not follow that. For example, the normalization is sometimes done by dividing the error indicators with the maximum-recorded value, resulting in a small nRMSE. Therefore, one needs to be cautious when citing and interpreting scale-independent errors reported in publications (Hoff et al., 2013).

On a separate note, the naming convention of the error indicators also varies largely in this category (see last column of Table 2). Unfortunately, it is difficult to define the "correct" naming convention.[12] Hence, the most common naming convention is selected here as the "correct" one. However, some incorrect usage, such as "mean average percentage error" (which, in our view, does not make any sense) still appears as many as four times in the retrieved 249 full texts.

Scale-independent error indicators can be used to compare forecast performance between different datasets. However, the accuracy of solar forecasting varies geographically due to differing weather and climate conditions and associated variability in solar irradiance. In addition, the forecast horizon and temporal resolution have a noticeable effect on forecast accuracy. To make forecast accuracies from different datasets more comparable, the forecast skill (FS) metric is often used, due to its popularity in weather forecasting. FS is computed by dividing the error indicator for a particular model (e.g., RMSE or MAE) with the

[11] In Table 4, ETS stands for ExponenTial Smoothing. This abbreviation is given by Hyndman et al. (2008). The abbreviation also refers to Seasonal, Trend and Error components of an ETS model.

[12] The usage of "deviation" instead of "error" has become more common in the recent literature. For now, the "error" terminology is adopted because it is still the primary choice in the literature, as shown in Fig. 5.

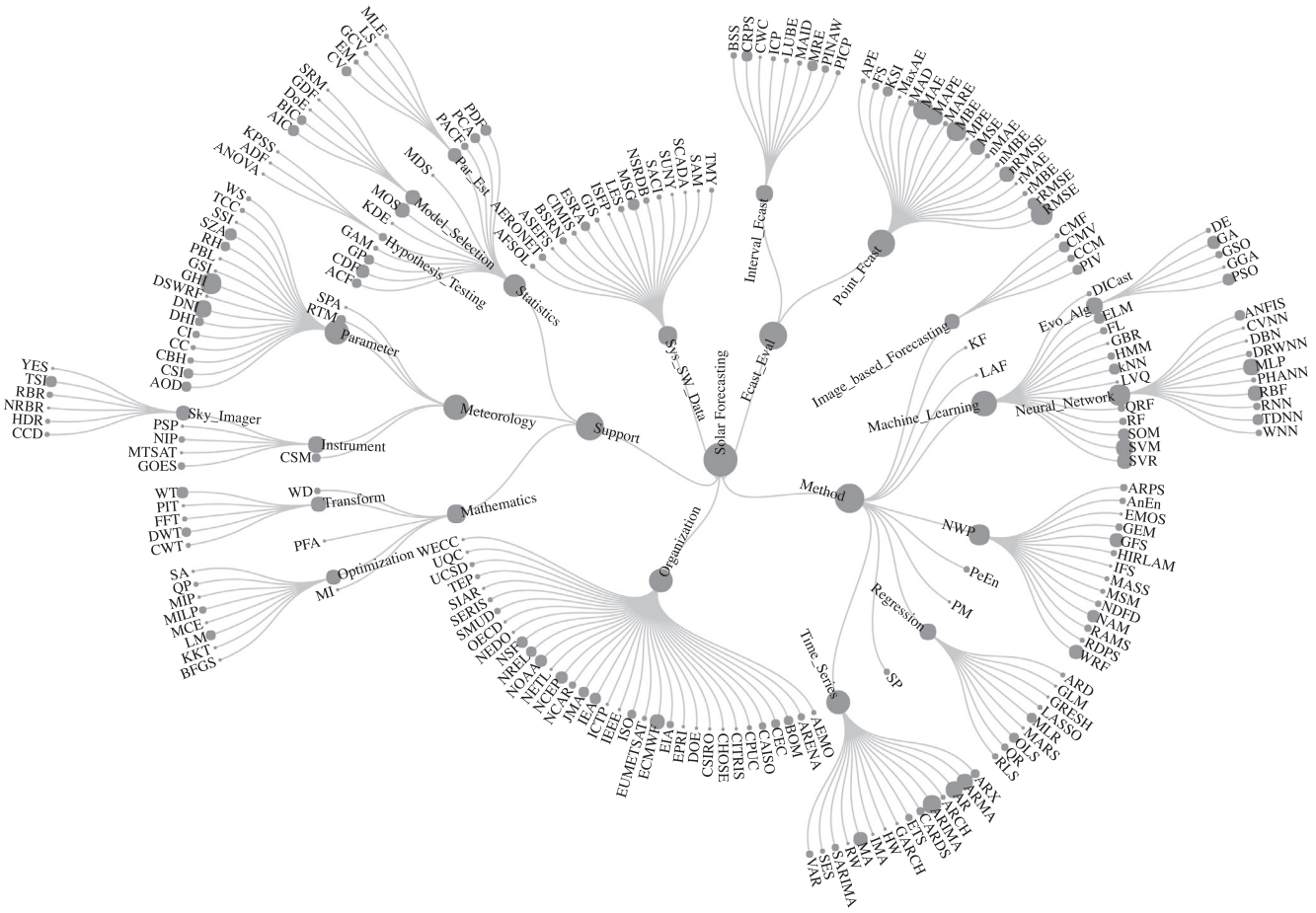

Fig. 5. Dendrogram of abbreviations with at least two citations in 249 solar forecasting papers on ScienceDirect. A version with both the abbreviations and the long forms is provided in the supplementary materials.

corresponding error indicator of a reference model (usually being the season-adjusted persistence model, see below). This fraction is then subtracted from 1, so that forecasts better than the reference model yield a positive skill. Due to its robustness and fairness, the FS originally proposed by Marquez and Coimbra (2012) has been recommended as one of the most suitable metrics for solar forecasting.

Another notable metric for point forecast evaluation is the Kolmogorov–Smirnov test integral (KSI). Instead of comparing each forecast with its respective measured value, KSI compares the forecast *distribution* to that of measurements. KSI also provides information on the variability of the forecast relative to the measurements; often a less variable forecast that tends towards the mean measurement yields a smaller RMSE, but may not be desirable from a user standpoint (Lorenzo et al., 2015). A less variable forecast could be detected by a larger KSI. To that end, Espinar et al. (2009) proposed KSI as an alternative metric to compare the distributions of time series predictions and observations.[13] Espinar et al. (2009)'s approach has gained acceptance in solar resource assessment applications (Gueymard, 2014). As shown in Table 2, KSI essentially evaluates the area between two empirical CDFs. The reader is referred to Huang and Thatcher (2017) and Perez et al. (2013, 2010) for example usages of KSI.

[13] The concept of KSI relies on the empirical cumulative distribution function (CDF), which puts a weight $1/n$ at each data point y_i, thus constituting an unbiased estimator of population CDF. Formally, the Kolmogorov–Smirnov test is used to test the hypothesis on equality of an empirical CDF with a parametric reference CDF (Massey, 1951). However, since the test itself only provides information on whether the two distributions are the same, it does not provide a quantitative measure of deviation in distributions over the entire range of the random variable. Furthermore, the reference CDF in a solar forecasting context is nonparametric, since it is the empirical CDF of measurements.

5.1.2. Error metrics for probabilistic forecast

Probabilistic forecasting aims either at providing a density function of the forecast value or at generating the prediction interval of the forecast (van der Meer et al., 2017). The error metrics used to evaluate these two types of probabilistic forecasting are also different. More specifically, the Brier skill score (BSS) and the continuous ranked probability score (CRPS) are used when the distribution of a forecast is available, whereas the remaining metrics shown in Table 3 are used to evaluate prediction intervals.

The Brier skill score is similar to the forecast skill: it compares the probabilistic forecast accuracy of a model—in terms of Brier score (Brier, 1950)—to that of a reference model. The Brier score (BS) can be considered the equivalent of the mean-squared error for probabilistic forecasts (Chu and Coimbra, 2017). BS is a verification tool for the prediction of the occurrence of an event. In its original form, BS—also called "probability score" according to Tödter and Ahrens (2012)—is evaluated by assigning probabilities (p_{ti}, $i = 1,...,m$) to m mutually exclusive categories. The sum of squared differences between p_{ti} and the true probability (1 if the event occurs in category i, 0 otherwise, represented by o_{ti}) is then computed, i.e., $\sum_{i=1}^{m}(p_{ti}-o_{ti})^2$. BS for n forecasts is therefore $(1/n)\sum_{t=1}^{n}\sum_{i=1}^{m}(p_{ti}-o_{ti})^2$. The reader is referred to the numerical example shown in the brief three-page paper by Brier (1950) for further clarification on probability score computation. Although this computation is capable of handling multicategory predictands, the probability score is known to be dependent on the choice of category threshold. It also disregards the distribution of probability mass with respect to its distance from the observed category (Tödter and Ahrens, 2012). Therefore, a better accepted version of BS is given by $(1/n)\sum_{t=1}^{n}(p_t-o_t)^2$, where p_t is the forecast probability of an event, and

Table 2
Frequently-used abbreviations in solar forecasting: forecast evaluation—point forecast. Forecast error at time t is denoted with $e_t = y_t - \hat{y}_t$, where y_t and $\hat{y}_t$ are actual and forecast values, respectively. This definition follows Hyndman et al. (2008).

Abbrv.	Long form	Computation	Remark	Confusion (C) or incorrect usage (I)
APE	Absolute Percentage Error	See MAPE	Absolute percentage error for a single point forecast, usually reported by taking average of n APEs	–
FS	Forecast Skill	$1-\frac{\text{nRMSE of your model}}{\text{nRMSE of a reference model}}$	A metric to compare a specific model to a reference model (usually persistence), independent of forecast horizon, location or method; it may be the most neutral and useful error metric in solar forecasting; nRMSE is sometimes replaced by nMAE and/or other statistical indicator	–
KSI	Kolmogorov Smirnov test Integral	$\int_{-\infty}^{\infty} \lvert F_n^y(x) - F_n^{\hat{y}}(x) \rvert \, dx$, where F_n^y and $F_n^{\hat{y}}$ are empirical CDFs of measurements and forecasts, respectively	Integrated differences (area) between the empirical cumulative distribution functions of measurements and forecasts; a discretized version of the integral is commonly used (see Espinar et al. (2009), for details)	Kolmogorov Smirnov Index (I); Kolmogorov Smirnov Integral (C)
MaxAE	Maximum Absolute Error	$\max \lvert e_t \rvert$	Evaluates the largest forecast error	–
MAD	Mean Absolute Deviation	See MAE	Same as MAE	Mean Absolute Distance (C)
MAE	Mean Absolute Error	$\frac{1}{n}\sum_{t=1}^{n} \lvert e_t \rvert$	Reflects the average magnitude of the errors	Mean Average Error (I)
MAPE	Mean Absolute Percentage Error	$\frac{100}{n}\sum_{t=1}^{n} \left\lvert \frac{e_t}{y_t} \right\rvert$	Scale-independent version of MAE, frequently-used to compare forecast performance between different data sets	Mean Average Percentage Error (I); Mean Absolute Percent Error (I)
MARE	Mean Absolute Relative Error	See rMAE	Same as rMAE	–
MBE	Mean Bias Error	$-\frac{1}{n}\sum_{t=1}^{n} e_t$	Evaluates forecast bias; the negative sign ensures a positive MBE corresponds to an overprediction.	Mean Biased Error (I)
MPE	Mean Percentage Error	–	Scaled version of MBE, not a standard error metric and not very useful	–
MSE	Mean Squared Error	$\frac{1}{n}\sum_{t=1}^{n} (e_t)^2$	Another scale-dependent measure, similar to MAE, but penalizes the larger errors	Mean Square Error (I)
nMAE	normalized Mean Absolute Error	$\left[\frac{1}{n}\sum_{t=1}^{n} \lvert e_t \rvert\right] \Big/ \left[\frac{1}{n}\sum_{t=1}^{n} y_t\right]$	MAE normalized by a factor, usually being the mean of measurements	–
nMBE	normalized Mean Bias Error	$\left[-\frac{1}{n}\sum_{t=1}^{n} e_t\right] \Big/ \left[\frac{1}{n}\sum_{t=1}^{n} y_t\right]$	MBE normalized by a factor, usually being the mean of measurements	–
nRMSE	normalized Root Mean Square Error	$\sqrt{\frac{1}{n}\sum_{t=1}^{n} (e_t)^2} \Big/ \left[\frac{1}{n}\sum_{t=1}^{n} y_t\right]$	RMSE normalized by a factor, usually being the mean of measurements	normalized Root Mean Squared error (I)
rMAE	relative Mean Absolute Error	See nMAE	Same as nMAE	–
rMBE	relative Mean Bias Error	See nMBE	Same as nMBE	–
rRMSE	relative Root Mean Square Error	See nRMSE	Same as nRMSE	–
RMSE	Root Mean Square Error	$\sqrt{\frac{1}{n}\sum_{t=1}^{n} (e_t)^2}$	The most well-applied error metric in solar forecasting; it penalizes the larger forecast errors	Root Mean Squared error (I); Root of Mean Square Error (I)

Table 3
Frequently-used abbreviations in solar forecasting: forecast evaluation—probabilistic forecast.

Abbrv.	Long form	Computation	Remark	Confusion (C) or incorrect usage (I)
BSS	Brier Skill Score	$1-\frac{\mathrm{BS}}{\mathrm{BS_{ref}}}$, where $\mathrm{BS}=\frac{1}{n}\sum_{t=1}^{n}\sum_{i=1}^{m}(p_{it}-o_{it})^2$ is the Brier score and $\mathrm{BS_{ref}}$ is the BS of a reference model; p_{it} is the probability that the forecast at time t falls in category i; and o_{it} takes the value 1 or 0 according to whether or not the event occurred in category i	A scoring rule to compare probabilistic forecast of a specific model to a reference model, while BS itself is considered as the mean square error for probabilistic forecasts	–
CRPS	Continuous Ranked Probability Score	$\frac{1}{n}\sum_{t=1}^{n}\int_{-\infty}^{\infty}(F^{\hat{y}_t}(x)-\mathbb{1}(x-y_t))^2dx$, where $F^{\hat{y}_t}$ is the CDF of the forecast $\hat{y}_t$ and $\mathbb{1}(x-y_t)$ is the Heaviside step function shifted to y_t	A scoring rule that generalizes mean absolute error, often used to evaluate meteorological forecasts	Continuous Rank Probability Score (I)
CWC	Coverage Width based Criterion	$\mathrm{PINAW}(1+\gamma(\mathrm{PICP})e^{-\eta(\mathrm{PICP}-\mu)})$, where μ is preassigned PICP to achieve, i.e., $(1-\alpha)$; η is the penalty strength; and $\gamma(\mathrm{PICP})=\begin{cases}0, & \text{if PICP}\geqslant\mu\\ 1, & \text{otherwise}\end{cases}$	A metric to assess PICP and PINAW simultaneously	Coverage Width Criterion (C)
ICP	Interval Coverage Probability	See PCIP	Same as PICP	–
MAID	Mean Absolute Interval Deviation	$\frac{1}{2n}\sum_{t=1}^{n}(\lvert L_t'-L_t\rvert+\lvert U_t'-U_t\rvert)$, where L_t and U_t are lower and upper bound of the prediction interval, L_t' and U_t' are lower and upper bound of the actual interval for time t	Measures the deviation of the predicted interval from the actual interval	–
MRE	Mean Relative Error	$\frac{100}{2n}\sum_{t=1}^{n}\left(\frac{\lvert U_t'-U_t\rvert}{R}+\frac{\lvert L_t'-L_t\rvert}{R}\right)$, where R is the range of target values	A normalized version of MAID using the range of target values	Missing Rate Error (C)
PINAW	Prediction Interval Normalized Average Width	$\frac{1}{nR}\sum_{t=1}^{n}(U_t-L_t)$	An interval forecast metric to assess the width of the prediction intervals	Prediction Interval Normalized Averaged Width (I)
PICP	Prediction Interval Coverage Probability	$\frac{1}{n}\sum_{t=1}^{n}\epsilon_t$, $\epsilon_t=\begin{cases}1, & \text{if } [L_t,U_t]\\ 0, & \text{otherwise}\end{cases}$	An interval forecast metric to indicate how often forecasts fall within prediction intervals	–

o_t is the actual outcome (1 if it really occurs and 0 otherwise). It is obvious that with this definition, the best score a forecast can achieve is 0 and the worst is 1. Finally, BSS is computed using $1-\mathrm{BS}/\mathrm{BS_{ref}}$.

Similar to BSS, CRPS compares the distribution of a forecast (in terms of CDF) to that of the corresponding measurement (Matheson and Winkler, 1976). Since the measurement is assumed to happen with a probability of 1 at y_t, its CDF is thus a Heaviside step function shifted y_t units to the right, i.e., $\mathbb{1}(x-y_t)$. For a probabilistic forecast, CRPS is therefore given by $\int_{-\infty}^{\infty}(F^{\hat{y}}(x)-\mathbb{1}(x-y_t))^2dx$. The CRPS for n forecasts can be obtained by averaging over these forecasts, as shown in Table 3. Unsurprisingly, these features made CRPS become the most popular metric for probabilistic forecasting evaluation. It is remarkable that CRPS and BS are closely related, and even equivalent under some specific conditions (see discussions by Tödter and Ahrens, 2012; Hersbach, 2000). In general, it is recommended to use CRPS for continuous predictands, such as solar irradiance, and BS for binary predictions, e.g., clear vs. cloudy state of the sky.

As compared to CRPS and BSS, the evaluation metrics for prediction intervals are straightforward. The prediction interval coverage probability (PICP) evaluates how often the forecasts fall within the prediction interval. In parallel, the prediction interval normalized average width (PINAW) assesses the width of the prediction interval. It is clear that these two metrics can be conflicting, i.e., both a high PICP and a narrow PINAW are desired. In the study by Quan et al. (2014), a metric called coverage width-based criterion (CWC) is used to combine the two metrics mentioned above. CWC penalizes the forecasts when the preassigned PICP (e.g., the nominal prediction interval) is not satisfied. This penalty forces the good forecasts to have a PICP close to the nominal prediction interval. Besides PINAW, it is also of interest to measure the deviation of the prediction interval from the actual interval. The metric called mean absolute interval deviation (MAID) examines the absolute errors between the predicted and actual interval's upper/lower bound pairs (Rana et al., 2015). The actual intervals are estimated using the next k data points counting from the current time step. A normalized version of MAID is also considered by Rana et al. (2015).

5.1.3. Error metrics for forecasts using sky-imaging cameras

The above error metric abbreviations resulted from the filtering process, i.e., they have at least one appearance. However, an examination of the complete list of abbreviations (including those with only one appearance) yields a third type of error metrics, namely, the error metrics for image-based forecasts. For example, three such abbreviations were found in Chu et al. (2015b): ramp detection index (RDI), false ramp index (FRI), and ramp magnitude forecast index (RMI). These abbreviations, however, only appear in Chu et al. (2015b), thus are not included in Fig. 5. This is not a structural deficiency of the abbreviation-based knowledge discovery elaborated here, because one can always look at the complete list of abbreviations or expand the number of papers to be searched.

Generally speaking, solar forecasting using sky imagery lacks the capability of producing probabilistic forecasts (van der Meer et al., 2017), unless it is combined with machine learning or statistical methods (e.g., Chu et al., 2015a). Unlike Chow et al. (2011), most investigators do not leverage the opportunity to produce spatial forecasts around the sky imager location, but rather produce point forecasts at the sky imager location. The metrics listed in Section 5.1.1 can also be used to evaluate the sky imager point forecast performance. However, due to the unique spatial and spectral nature of sky imagery, many method-specific metrics have been developed or utilized. For instance, precision, recall, and F_2 score are used to evaluate the accuracy of cloud occlusion prediction (Bernecker et al., 2014). Similarly, the "accuracies of cloud" metric is used to measure the error in cloud classification (Peng et al., 2015), and the cloud-advection-versus-persistence (cap) error—a metric that is the equivalent to forecast skill—is used to describe the forecast error obtained by cloud advection (Chow et al.,

Table 4
Frequently-used abbreviations in solar forecasting: method—time series.

Abbrv.	Long form	One-sentence description	Further reading	Confusion (C) or incorrect usage (I)
ARX	AutoRegressive with eXogenous input	An autoregressive model with additional regressors (exogenous inputs), such as cloud cover estimates from an NWP model.	Bacher et al. (2009)	AutoRegressive model with eXogenous input (C); Auto Regressive eXogenous (C); AutoRegressive with eXogenous (C)
ARMA	AutoRegressive Moving Average	A time series modeling approach that considers both AR and MA processes, for stationary time series.	See further reading for ARIMA	AutoRegressive and Moving Average (I)
AR	AutoRegressive	The current value is modeled as a function of p past values.	See further reading for ARIMA	Auto-Regressive (I); Autoregression (C); Auto Regressive (I)
ARCH	AutoRegressive Conditional Heteroscedastic	A stochastic process that performs heteroscedasticity corrections by modeling the conditional variance of the current error term.	Boland (2015), Engle (1982)	–
ARIMA	AutoRegressive Integrated Moving Average	The generalization of ARMA modeling that includes differencing; it is able to handle nonstationary time series.	Reikard (2009), Box and Jenkins (1994)	–
CARDS	Coupled AutoRegressive and Dynamical System	A hybrid model that combines an AR model and a dynamical system model, designed specifically for solar forecasting.	Huang et al. (2013)	
ETS	ExponenTial Smoothing	A state space modeling framework incorporating 30 stochastic models, likelihood calculation, prediction intervals and model selection.	Yang et al. (2015c), Dong et al. (2013), Hyndman et al. (2008)	Exponential Smoothing State Space (I)
GARCH	Generalized AutoRegressive Conditional Heteroskedasticity	GARCH models the error variance of an autoregressive model with an ARIMA model.	David et al. (2016), Sun et al. (2015), Bollerslev (1986)	–
HW	Holt Winter	An ETS method for dealing with data that contains both a linear trend and a seasonal component.	See further reading for ETS	–
IMA	Integrated Moving Average	A variant of ARIMA with an AR process order of $p = 0$.	See further reading for ARIMA	–
MA	Moving Average	The current value is modeled as a function of current and q past values of a white noise process.	See further reading for ARIMA	–
RW	Random Walk	The current value is modeled as the previous value plus a white noise term.	See further reading for ARIMA	
SARIMA	Seasonal AutoRegressive Integrated Moving Average	A model formed by including additional seasonal terms in the ARIMA models.	Aryaputera et al. (2015a), Bouzerdoum et al. (2013), and Box and Jenkins (1994)	–
SES	Simple Exponential Smoothing	An ETS method for nonseasonal data without trend.	See further reading for ETS	–
VAR	Vector AutoRegressive	A generalization of AR models that considers linear interdependencies across multiple series, e.g., sensor network and satellite-derived irradiance series.	Bessa et al. (2015) and Yang et al. (2014a)	Vector AutoRegression (C); VARiational (C)

Table 5
Frequently-used abbreviations in solar forecasting: method—regression.

Abbrv.	Long form	One-sentence description	Further reading	Confusion (C) or incorrect usage (I)
ARD	Automatic Relevance Determination	A technique under Bayesian framework to assign regularization coefficients (weights) to input variables, used in both neural network training and linear regression fitting	Mazorra Aguiar et al. (2016) and Mellit and Pavan (2010)	–
GLM	Generalized Linear Model	A linear regression model that allows the response to have a non-normal distribution, e.g., logistic regression and Poisson regression are GLM.	Voyant et al. (2017c)	–
GRESH	Group Regularized Estimation under Structural Hierarchy	A very recent and advanced hierarchical variable selection method published in JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION that considers both LASSO penalty and group LASSO penalty.	Jiang et al. (2017) and Jiang and Dong (2017)	–
LASSO	Least Absolute Shrinkage and Selection Operator	A correlation-based regression predictor selection method with $\mathscr{L}_1$-regularization, primarily used to select spatio-temporal neighbors in a sensor network	Jiang et al. (2017) and Yang et al. (2015d)	–
MLR	Multiple Linear Regression	Linear regression with two or more predictor variables, e.g., location, time, or meteorological variables.	Deo and Şahin (2017) and Wang et al. (2016b)	Multivariate Linear Regression (C)
MARS	Multivariate Adaptive Regression Spline	A non-parametric method for flexible modeling of high-dimensional data; it partitions the input space into regions, each with its own regression equation.	Massidda and Marrocu (2017) and Li et al. (2016c)	–
OLS	Ordinary Least Squares	A method to estimate parameters in a linear regression model, by minimizing the residual sum of squares.	Yang et al. (2017b), Brabec et al. (2015), and Jiang et al. (2015)	Ordinary Least Square (I)
QR	Quantile Regression	A regression method that estimates the conditional median (rather than the conditional mean, as in OLS) or other quantiles of the response variable, by minimizing asymmetric penalties that are functions of quantiles and least absolute deviations.	Brabec et al. (2015) and Zamo et al. (2014a)	–
RLS	Recursive Least Squares	A variant of LS method that minimizes weighted least squares, suitable for online short-term forecasting.	David et al. (2016) and Bacher et al. (2009)	Recursive Least Square (I)

2011). Some metrics are similar or equivalent to each other, but named differently. A more thorough review of sky-imagery-based solar forecasting literature is needed to compare and discuss various metrics and converge on a recommended suite of metrics for assessing the performance of sky-imagery-based forecasting.

5.2. Solar forecasting method

In the review paper by Inman et al. (2013), one of the earliest and perhaps the most well-accepted classification of solar forecasting methods is presented. A total of five classes of methods, namely, wireless sensor network, total sky images, satellite imaging, NWP and stochastic & artificial intelligence, are compared based on their spatial resolution and typical forecast horizon (see Fig. 20 in Inman et al., 2013). Since then, several similar classifications and plots have been presented, e.g., see Figs. 8 and 4 in Antonanzas et al. (2016) and Diagne et al. (2013), respectively. In this section, solar forecasting methods are classifed into five classes, based on the list of frequently-used abbreviations: (1) time series; (2) regression; (3) numerical weather prediction; (4) machine learning; and (5) image-based forecasting. The abbreviations in each class are summarized in Tables 4–8. To avoid repetition of previous reviews, instead of explicitly explaining each method, only a one-sentence description is given in these tables. However, the most representative references are listed as further reading material.

5.2.1. Time series

A method is classified as a time series method if it falls in one of three families of models, namely, autoregressive integrated moving average (ARIMA), exponential smoothing (ETS), and generalized autoregressive conditional heteroskedasticity (GARCH). Largely owing to the two books written by Box and Jenkins (1994) and Hyndman et al. (2008), the ARIMA and ETS families of models became iconic for the subject of statistical forecasting. The GARCH family of models—first proposed by Engle (1982) and then generalized by Bollerslev (1986)—gained its recognition in econometrics due to its ability to perform heteroscedasticity corrections; it is widely used to model financial time series. For the needed complete review on time series forecasting methods, the reader is referred to the review by Gooijer and Hyndman (2006).

In terms of adaptation to meet solar forecasting needs, Fig. 5 reveals that ARIMA is the most-widely-used time series method, essentially because it is a common choice for a reference method. Although studies have shown that ETS and GARCH outperform ARIMA (e.g., Dong et al., 2013), it is advised to keep ARIMA as a candidate/reference model during forecasting, due to its enormous impact on the theory and practice of modern time series analysis and forecasting. Furthermore, many statistical software packages, typified by the popular **forecast** package in **R**, provide automatic model identification and parameter estimation capabilities for ARIMA, and thus keep the effort of solar forecast practitioners to a minimum.

The main drawback of time series models in solar forecasting is perhaps the lack of physical modeling during forecasting. The variability of solar irradiance and related time series is mainly due to moving clouds and weather systems. Ignoring these factors in time series forecasting often results in forecasts that appear to be lagging. On this point, the multivariate versions of time series models, such as the autoregressive with exogenous input (ARX) and vector autoregressive (VAR), may produce improved forecasts. Nevertheless, the amount of improvement depends on one's understanding on the predictors, and how they affect the predictand. Adding irrelevant predictors, or adding inadequate forms of them, into a multivariate model contributes to the overall prediction variance.

5.2.2. Regression

Regression is a statistical process for estimating the relationships

Table 6
Frequently-used abbreviations in solar forecasting: method—NWP.

Abbrv.	Long form	One-sentence description	Further reading	Confusion (C) or incorrect usage (I)
ARPS	Advanced Regional Prediction System	An NWP model for mesoscale simulation and forecasts developed by the University of Oklahoma http://www.caps.ou.edu/ARPS/	Law et al. (2014) and Perez et al. (2013)	–
AnEn	Analog Ensemble	A method to predict the probability distribution of a future state of the atmosphere using past predictions (analogs) of a deterministic NWP model	Cervone et al. (2017) and Alessandrini et al. (2015)	–
EMOS	Ensemble Model Output Statistic	A statistical postprocessing technique that addresses the systematic biases and underdispersive ensembles	Sperati et al. (2016) and Gneiting et al. (2005)	–
GEM	Global Environmental Multiscale	The integrated forecasting and data assimilation system developed by Environment Canada http://collaboration.cmc.ec.gc.ca/science/rpn/gef_html_public/index.html	Larson et al. (2016) and Pelland et al. (2013)	Global Environmental Multiscale model (C)
GFS	Global Forecast System	A global weather forecast model produced by NCEP http://www.nco.ncep.noaa.gov/pmb/products/gfs/	Gagne et al. (2017) and Verzijlbergh et al. (2015)	Global Forecasting service (I); Global Forecasting System (I)
HIRLAM	HIgh Resolution Limited Area Model	The NWP system developed by the international HIRLAM programme http://hirlam.org/	Larson et al. (2016) and Perez et al. (2013)	HIgh Resolution Limited Area Modelling (I)
IFS	Integrated Forecasting System	The atmospheric model and data assimilation system developed by European Centre for Medium-Range Weather Forecasts	ECMWF (2017) and Morcrette et al. (2008)	Integrated Forecast System (I)
MASS	Mesoscale Atmospheric Simulation System	An NWP model for mesoscale simulation and forecasts	Perez et al. (2013) and Manobianco et al. (1996)	Mesoscale Atmospheric Simulation System model (C)
MSM	Mesoscale Model	Although the name may be confusing, this acronym usually refers to the NWP model developed by JMA	Ohtake et al. (2015) and Ohtake et al. (2013)	Meso-scale Model (I)
NDFD	National Digital Forecast Database	A suite of gridded weather forecast products https://www.weather.gov/mdl/ndfd_home	Perez et al. (2013) and Marquez and Coimbra (2011)	National Digital Forecasting Database (I)
NAM	North American Mesoscale	A regional weather forecast models run operationally by NCEP http://www.nco.ncep.noaa.gov/pmb/products/nam/	Larson et al. (2016) and Mathiesen and Kleissl (2011)	North American Model (I); North American Mesoscale model (C)
RAMS	Regional Atmospheric Modeling System	An NWP model for mesoscale simulation and forecasts	Alessandrini et al. (2015) and Pielke et al. (1992)	–
RDPS	Regional Deterministic Prediction System	An NWP model developed by the Canadian Meteorological Centre; often used by researchers for its cloud cover data	Nonnenmacher et al. (2016) and Larson et al. (2016)	–
WRF	Weather Research and Forecasting	An NWP model for mesoscale simulaton and forecasts jointly developed by NCAR and NOAA; a specific configuration of WRF, namely, **WRF-Solar** (Jimenez et al., 2016) is the first NWP model specifically designed for solar power prediction	Yang and Kleissl (2016) and Lima et al. (2016)	Weather and Research Forecasting (I); Weather Research and Forecast (I); Weather Research Forecast (I)

Table 7
Frequently-used abbreviations in solar forecasting: method—machine learning.

Abbrv.	Long form	One-sentence description	Further reading	Confusion (C) or incorrect usage (I)
DICast	Dynamic Integrated foreCast	A consensus learning system—for weather variables—that can produce an optimized forecast based on a variety of input data, e.g., NWP outputs and ground-based measurements; its forecasts can be further nested as exogenous inputs for solar forecasting models	McCandless et al. (2016) and Mahoney et al. (2012)	–
ELM	Extreme Learning Machine	A very fast training method for feedforward neural networks; its output weights can be learnt in one step, by randomly assigning the input weights and computing the Moore–Penrose pseudoinverse of the hidden layer output matrix	Bouzgou and Gueymard (2017) and Salcedo-Sanz et al. (2014)	Extreme Learning Machine algorithm (C)
FL	Fuzzy Logic	A form of many-valued logic—in contrast with Boolean logic—with truth values of vaiables ranging from 0–1; it handles forecasting problems with, e.g., imprecise, uncertain, or nonlinear data	Chen et al. (2013) and Boata and Gravila (2012)	–
GBR	Gradient Boosted Regression	Staged (progressively optimized) predictive method based on classification and iterative improvement over ensembles of less accurate models	Persson et al. (2017) and Gala et al. (2016)	–
HMM	Hidden Markov Model	A tool for representing probability distributions over a sequence of observations generated by a hidden process that satisfies Markov properties; it is suitable for time series data modeling	Li et al. (2016a) and Soubdhan et al. (2016)	–
kNN	k-Nearest Neighbor	A non-parametric method used for classification and regression where the output depends on the k-th closest training samples of a look-up library of patterns	Chu and Coimbra (2017) and Pedro and Coimbra (2012)	k Nearest Neighbour (C); k Nearest Neighbor algorithm (C)
QRF	Quantile Regression Forest	A generalized form of the RF method used to predict conditional quantiles and other distribution properties	Almeida et al. (2017) and Zamo et al. (2014b)	–
RF	Random Forest	Ensemble learning methods based on decision trees and randomized feature selection	Urraca et al. (2016) and Almeida et al. (2015)	–
SVM	Support Vector Machine	Nonlinear classification or regression algorithms based on supervised mapping of categories on a set of hyperplanes that can be clustered by magnitude of a chosen norm	Jiang and Dong (2017) and Wang et al. (2015)	State Vector Machine (C); Support Vector Machines method (C)
SVR	Support Vector Regression	The SVM method adapted for regression by weighed consideration of all data points in the training set	Lin and Pai (2016) and Wolff et al. (2016)	Support Vector machine Regression (I)
Evolution algorithm				
DE	Differential Evolution	A simple and efficient adaptive scheme for global optimization over continuous spaces, one of the most powerful evolutionary algorithms for real number function optimization	Storn and Price (1997) and Jiang et al. (2015)	Diesel Engine (C)
GA	Genetic Algorithm	An evolutionary algorithm with binary-valued representations (solutions); it is based on the classic view of a chromosome as a string of genes	Zagouras et al. (2015) and Pedro and Coimbra (2012)	–
GSO	Glowworm Swarm Optimization	A swarm intelligence algorithm mimicking how glowworms use signaling and attraction mechanisms to congregate into large swarms; it handles problems with multiple optima of multimodal functions	Jiang et al. (2017) and Jiang and Dong (2016)	Genetical Swarm Optimization (C)
GGA	Grouping Genetic Algorithm	A kind of GA, modified to suit the structure of grouping problems; it was used to select WRF outputs (features) as inputs to ELM for GHI prediction	Aybar-Ruiz et al. (2016)	–
PSO	Particle Swarm Optimization	A robust stochastic optimization technique based on the movement and intelligence of a number of agents (particles) that constitute a swarm	Ni et al. (2017) and Dong et al. (2015)	–
Artificial neural network				
ANFIS	Adaptive Neuro Fuzzy Inference System	A combination of ANN and fuzzy logic; for instance, ANN can be used to adjust the membership functions in fuzzy logic	Bigdeli et al. (2017) and Sfetsos and Coonick (2000)	Artificial Neuro Fuzzy Inference System (I)
DBN	Deep Belief Network	A multilayer generative model where each layer encodes statistical dependencies among the units in the layer below	Hinton et al. (2006) and Dedinec et al. (2016)	Dynamic Bayesian network (C)
DRWNN	Diagonal Recurrent Wavelet Neural Network	A network combining RNN and WNN to benefit from both the dynamic properties of RNN and abilities of WNN to map nonlinear functions	Cao and Lin (2008)	–
LVQ	Learning Vector Quantization	A supervised classification algorithm that uses a winner-take-all Hebbian learning-based approach	Yang et al. (2014e)	Learning Vector Quantity (I)
MLP	MultiLayer Perceptron	The most fundamental NN structure that contains one input layer, one or more hidden layers, and one output layer	Voyant et al. (2017b) and Mellit and Pavan (2010)	Multi layer perceptron (I)
PHANN	Physical Hybrid Artificial Neural Network	An MLP that uses a physical clear-sky model as part of its inputs	Ogliari et al. (2017) and Antonanzas et al. (2016)	–
RBF	Radial Basis Function	A function which depends only on the radial distance from the input to a given center, herein refers to an NN structure that uses RBFs as activation function	Jiang and Dong (2016) and Jiang et al. (2015)	Radical Basis Function (I); Radical Basic Function (I)
RNN	Recurrent Neural Network	A class of ANNs whose connections between nodes form a loop; e.g., recursive NN, Hopfield network, or Elman network	Mellit et al. (2014) and Mellit and Pavan (2010)	Recursive Neural Network (C)
SOM	Self Organizing Map	A method to transform input patterns into a two-dimensional (2D) map of features; the data distribution becomes more uniform than the whole input data space, thus facilitating non-linear input–output mapping	Dong et al. (2015) and Ghayekhloo et al. (2015)	Self Organized Map (I)

(*continued on next page*)

Table 7 (*continued*)

Abbrv.	Long form	One-sentence description	Further reading	Confusion (C) or incorrect usage (I)
TDNN	Time Delay Neural Network	A class of ANNs for sequential data; the inputs to a node can consist of outputs of earlier nodes from current and previous time steps	Fernandez-Jimenez et al. (2012) and Wu and Chee (2011)	Time Delayed Neural Network (I)
WNN	Wavelet Neural Network	An ANN structure that uses wavelet functions as activation function for hidden neurons	Alexandridis and Zapranis (2013) and Sharma et al. (2016)	–

among variables; it can be very useful in solar forecasting in terms of modeling exogenous variables. The general form of a regression model, for a single predictand case, is $\boldsymbol{Y} = f(\boldsymbol{X},\boldsymbol{\beta})$, where $\boldsymbol{Y}$ is the vector of dependent variable, $\boldsymbol{X}$ is the matrix of independent variables, $\boldsymbol{\beta}$ is the regression parameter(s). The function f that relates $\boldsymbol{Y}$ to $\boldsymbol{X}$ can be linear or nonlinear. The forecasting methods discussed in this section are mostly linear or piecewise linear; the nonlinear case will be discussed in the neural network section below.

The (arguably) simplest regression model is a first-order autoregressive process, noted AR(1), i.e., $y_t = \beta_0 + \beta_1 y_{t-1} + e_t$, where β_0 is a constant and e_t is white noise. In this model, the only predictor is the lag-1 time series of the predictand, i.e., $\{y_{t-1}\}$. If more predictors are added to the model, the regression model is then known as a multiple linear regression (MLR) model. A naive way of constructing an MLR model is to include as many predictors as possible—wind speed, temperature, humidity, cloud cover, location, time and many other variables—that are thought to be relevant to solar irradiance into the model. Once the model is set, the regression parameters $\boldsymbol{\beta}$ can be estimated via the ordinary least squares (OLS) method. However, this is far from being satisfactory, if not completely useless, for two main reasons: (1) the OLS estimates have low bias but high variance, which affects the overall prediction accuracy when the number of predictors is large, and (2) the model is not interpretable (Hastie et al., 2009; Efron et al., 2004). For such reasons, variable selection is critical to improve predicability and interpretability. In fact, the variable selection problem is so important that just one of the representative papers, by Tibshirani (1996), has received more than 20,000 citations alone. The method proposed therein, namely, the least absolute shrinkage and selection operator (LASSO), has also been used in solar forecasting (Jiang et al., 2017; Yang et al., 2015d), alongside other variable selection methods, such as the automatic relevance determination (ARD) and group regularized estimation under structural hierarchy (GRESH).

A particularly promising application of MLR in solar forecasting is spatio-temporal forecasting using data from a monitoring network of appropriate size. As mentioned earlier, moving clouds are the main source of irradiance variability. As clouds propagate over the monitoring network, data collected by the neighboring sensors can be used as predictors for the forecast location. However, the number of predictors can become very large as the number of sensors in the network increases. In this situation, one can filter the predictors by only selecting the lagged time series collected by the upwind sensors. It has been shown that, by doing so, the selected upwind predictors can effectively help predict the ramp events at downwind locations and achieve high forecast skill (up to 0.5 for the networks used in Yang et al., 2015d). However, forecast horizons using these methods are often limited by the size of the monitoring network. The correspondence between spatial scale of the network and forecast horizon requires further study. The reader is referred to the studies by Aryaputera et al. (2015b), Lonij et al. (2013) and Yang et al. (2013b), as well as the review by André et al. (2016), for more discussion on the connections between spatio-temporal statistical models and regression-based forecasting. Redesigning existing monitoring networks for solar forecasting purposes was proposed in, e.g., Yang (2017).

5.2.3. Numerical weather prediction

Owing to the large heating effects of solar radiation throughout the atmosphere and at the surface, and to the resulting atmospheric circulation, all numerical weather prediction (NWP) models directly simulate the irradiance fluxes at multiple levels in the atmosphere, separately considering the shortwave and longwave parts of the solar spectrum. In other words, GHI (and to a lesser extent DNI) are prognostic variables in NWP. Until recently, however, the surface irradiance was usually not provided as a model output because it was deemed unimportant relative to temperature, wind, humidity, or precipitation. This was the case with the U.S. National Digital Forecast Database, for instance. The rising importance of solar irradiance for operating the

Table 8
Frequently-used abbreviations in solar forecasting: method—image-based forecasting.

Abbrv.	Long form	One-sentence description	Further reading
CMF	Cloud Motion Forecast	Despite the literal meaning of this abbreviation, it refers to a method that forecasts cloud cover, and thus clear-sky index, by translating a "frozen" (in the image sense, not as the state of water) cloud field	Hammer et al. (2003) and Marquez et al. (2013)
CMV	Cloud Motion Vector	Cloud motion vector field derived from consecutive satellite images, sky imagers or interpolated irradiance maps	Chow et al. (2015) and Nonnenmacher and Coimbra (2014)
CCM	Cross Correlation Method	A method to generate CMV by matching a block of pixels to candidate pixel blocks in the subsequent image; a vector is drawn from between the original block center and the center of the block with maximum correlation	Hamill and Nehrkorn (1993) and Chow et al. (2011)
PIV	Particle Image Velocimetry	A method to measure velocities in fluids; in solar forecasting, it is used to derived the average cloud velocity through two consecutive images	Li et al. (2016b) and Chu et al. (2015b)

electric power system has changed this; anyone with basic computer literacy can now download solar forecasts (of GHI only, most usually) up to multiple days ahead from national weather centers.

A multitude of NWP models exist (see Table 6 for those that are cited most frequently). Most, if not all, solve the fundamental equations of fluid motion, but their numerical scheme and models for subgrid-scale physical processes may differ. These subgrid-scale models or "physics options" are usually specific to a physical process: e.g. a planetary boundary layer (PBL) represents the unresolved effects of turbulent mixing in the lowest layer of the atmosphere. In WRF, for instance, one of 12 possible PBL options has to be selected.[14] Similarly, cloud microphysics, cumulus parameterization, shortwave and longwave radiation, and land surface models need to be selected, thus resulting in thousands of possible model configurations. Typically, only experts can understand the differences between physics options within a group. There are also interaction effects between different schemes, which can only be appreciated by the most experienced modelers. Therefore the learning curve for custom NWP forecasts is extremely steep. For an accessible review of the many challenges of simulating solar radiation with NWP models, see Larson (2013).

Some NWP models are popular because they are open-source and can be configured by the user to high resolution over a specific region, such as in WRF (Yang and Kleissl, 2016). Others are popular because of free access and global coverage, such as GFS. The ECMWF model has been shown to perform better than other forecast models, but is only available with a subscription.

Whereas originally a single deterministic model constituted the forecast, increases in computational power now allow running an ensemble forecast system of multiple model instances with different initial conditions, initialization times, and/or physics parameters (e.g., ECMWF Ensemble Prediction System or NOAA Global Ensemble Forecast System). Although they are typically underdispersive, ensemble forecasts (at least in theory) can be used to generate probabilistic forecasts. An alternative approach to probabilistic forecasting with NWP is to utilize a long time history of the predictor and predictand from a single model together with statistical approaches (e.g., the Analog Ensemble method of Alessandrini et al., 2015). For details on NWP models, the reader is referred to Chapters 12–14 of Kleissl (2013).

5.2.4. Machine learning

Nowadays, machine learning (ML) is perhaps the most popular approach in solar forecasting. This is partly owing to the large number of available methods and variants in ML, as shown in Fig. 5, as well as the wide range of applications it supports, including classification, regression, and clustering. ML algorithms have a long history, with key foundational events occurring in the middle of the 20th century. The simplest learning machine is Rosenblatt's perceptron (Rosenblatt, 1958; McCulloch and Pitts, 1943), developed during the formative years of artificial neural networks (ANNs)—early 1940s to late 1950s. It consists of a linear combiner and a hard limiter, so that the perceptron produces a "+1" or a "−1" depending on the inputs. It is thus a classifier. Despite the major setback in the 1970s, largely owing to the criticism of perceptrons made by Minsky and Papert (1969), the work of Rosenblatt was eventually incorporated in the more general framework of multilayer perceptrons (MLPs) with backpropagation (BP) (Werbos, 1974), one of the most popular ANN architectures due to its ability to perform arbitrary non-linear mappings. Since those days, ML has been greatly expanded, and ML research is going unprecedentedly strong.

Despite their diverse nature, ML algorithms share the same framework: they are based on the concept of learning patterns and model parameters from the data, where learning implies classification, regression, and prediction. In this sense, ML algorithms are well suited for solar forecasting. Solar forecasting applications consist of creating predictive models for point values or prediction intervals based on a dataset of historical data. The dataset typically contains the target or endogenous variable (irradiance or PV power output) and may contain any number of exogenous predictors such as NWP forecasts or meteorological data. Interested readers are referred to Voyant et al. (2017c) for a review on ML methods for solar forecasting.

Solar forecasting publications based on ANNs can be classified into five major types (not mutually exclusive):

1. Hybrid methods—as mentioned earlier, combining ANN with other methods leads to likely improvements in forecast accuracies.
2. Alternatives to conventional statistical methods—ANN is capable of performing regression tasks. It is thus often used to replace the methods discussed in Section 5.2.2.
3. Applying different ANN structures to solar forecasting—many ANN structures[15] are developed in the ML community. Some structures are being transferred to solar forecasting applications.
4. Location-specific validation reports—during the early years of solar forecasting, many publications reported the accuracies of ANN based on local data. However, this type of studies is slowly being phased out, since the current trend is to obtain universally applicable results.
5. Comparison papers and reviews.

Although ANNs and support vector machines (SVMs) remain popular as the basis for ML methods in solar forecasting, many other approaches, such as k-nearest neighbors (kNN), random forest (RF) or gradient boosted regression (GBR), have been used lately (see Table 7 for more details and references). For a given set of input data, the proper implementation of any of these methods yields similar forecasting skills, as long as overfitting is prevented.

Whereas many ways exist to identify, classify and predict patterns from data, the common characteristic of most ML methods is to train, test, validate, and verify (using some error metrics) against subsets of

[14] http://www2.mmm.ucar.edu/wrf/users/phys_references.html.

[15] Commonly used ANN structures include MLP, radial basis function (RBF), self organizing map (SOM), etc.

the available historical data, in order to prevent overfitting. In their simplest forms, ML methods are statistical methods capable of identifying trends and substantially reducing bias with respect to the validation set. The robustness of ML models depends on the diversity of the training data set, the training method, the ability of the endogenous and exogenous variables to capture the space of solutions required to reproduce the outputs, and very importantly, the figure of merit used to validate the results. Overlaying ML methods with other methods almost always results in better forecasts if a broad set of forecasting accuracy metrics is used (exemplified by the many works of the Coimbra Research Group, e.g., Chu et al., 2016; Nonnenmacher et al., 2016; Chu et al., 2015a,b, 2013; Marquez et al., 2013; Pedro and Coimbra, 2012).

While excelling in predictive power, as demonstrated in the publications listed in Table 7, the black-box nature of some of the ML algorithms—ANNs most notably—provides very little insight, if not no insight at all, into the underlying physical relationships between inputs and outputs (Brabec et al., 2015). This situation is not irremediable since there are several tools to explore and understand the mechanics of ANNs (e.g., neural interpretation diagram, Garson's algorithm and sensitivity analysis; Olden et al., 2004). However, in general, such studies are not carried out in the solar forecasting domain because the forecasting performance is more valued than the model's explanatory power.

Key model settings, such as the input selection and the topology of the ML non-linear approximators, are often subject to an optimization procedure at the training stage. This results in more complex renditions, in which the ML methods employ master evolutionary algorithms, such as genetic algorithm (GA) and particle swarm optimization (PSO), to dynamically optimize the topology of nonlinear approximators, e.g., ANNs. In this case, the input selection methodology and the model features themselves are optimized stochastically as new data becomes available. Hence, the model information stored in the topology of the networks is adaptively evolving in response to pattern changes in the input data.

Because the accuracy and robustness of the forecasts produced by ML methods depend on both the training method and the figure of merit used to evaluate the quality of the forecasts, those two components of the methodology require special attention. An arbitrary partition of the data into training, verification and validation sets is no longer considered ideal (see Chu et al. (2013) for a discussion on cross validation versus randomized training and validation). With increased computational power and with enough historical data available, each set is better determined through a multi-objective optimization using several figures of merit to avoid MBE or RMSE reductions at the expense of ramp capturing or variability smoothing. The cascading processes that determine optimal recursive strategies for data partition, input selection, training, verification, and validation through convergence techniques ultimately result in much less need for input from the modeler, since multiple steps along the process are determined exclusively by the available data. The combination of the cascading optimization with hierarchical classes of representation of data leads to the concept of *deep learning*.

5.2.5. Image-based forecasting

Sky or earth imagery can add predictive skill because it provides advance warning of approaching clouds at a lead time of several minutes to hours. This lead time far exceeds that of a single ground-based radiometer. With the exception of thick overcast conditions, GHI (as measured with, e.g., a pyranometer) is dominated by the state of the atmosphere (cloudy versus clear) along the sun-instrument slant path. Approaching clouds therefore largely remain unobservable—an exception being caused by cloud enhancement (Pecenak et al., 2016)—until the cloud boundary has started to block the sun. A complete occlusion of the solar disk then occurs within seconds or less, which is too little advance warning for most applications. Many sky-imager forecasting approaches leverage the spatial nature of imaging data only for model training, whereas forecasts are provided as point forecasts for the sky imager location only. Such approaches typically extract image data only along a line or a sector upwind of the sun, and apply regression and/or machine learning methods to imager measurements at specific pixels or groups of pixels to derive GHI and/or DNI at the sky imager location. Such approaches cannot necessarily be applied to other pixels in the image, since fisheye lenses cause the projection of cloud motion to become increasingly non-linear away from the position of the sun.

This brief review of image-based forecasting mostly applies to satellite imagery and a few spatial sky-imager forecasting methods (e.g., Chow et al., 2011). If the 3D nature of clouds and the cloud height are ignored, as in most satellite imaging approaches (see Miller et al. (2018), for a discussion of impacts of sun-ground-satellite geometry effects), the geometry of the problem becomes trivial: the reflectance enhancement in a satellite pixel compared to the clear-sky's background reflectance is associated with clouds over the underlying ground pixel (Table 10).

The remaining tasks in solar forecasting then consist of determining (i) the GHI and/or DNI fields at the surface at the time of the image, and (ii) the cloud motion to advect those fields into the future. For a review of satellite-based solar resource models, see Chapters 2 and 3 in Kleissl (2013). The clear-sky index is often calibrated against the relative reflectance enhancement, as for example in the Perez model (Perez et al., 2002), which continues to be used widely with only minor refinements to this day. The cloud motion vector field—as determined for example using the block matching (e.g., Yang et al., 2013c; Li et al., 1994) or cross-correlation method (CCM), or particle image velocimetry (PIV)—is applied to advect the current GHI field into the future.

To increase the spatial resolution of forecasts based on sky imagers, the detailed sun-clouds geometry with respect to the ground is of the greatest importance. The simple geometry in satellite image-based forecasting is therefore advanced to allow the clouds to exist in an infinitesimally thin layer of the atmosphere—the cloud base height (Chow et al., 2011). However, Kurtz et al. (2017) show that, even then, perspective issues contribute to the majority of forecast errors when a single imager is used and the 3D nature of clouds is neglected. Accurate spatial sky-imager forecasting requires multiple distributed sky cameras to resolve the 3-dimensionality of clouds. Such work has just started recently (Peng et al., 2015). In addition to the adverse perspective of sky images, the solar resource analysis faces significant challenges: (i) the sky-imager radiometry being challenging (Kurtz and Kleissl, 2017), spectral information (e.g., the red-to-blue-ratio, Table 10) is often rather used to infer clouds and cloud optical depth, but this provides non-unique solutions (Mejia et al., 2015); (ii) the most critical image area near the sun is difficult to observe due to signal saturation; and (iii) soiling seriously affects the quality of sky images, unless the camera is equipped with a good ventilator. The authors believe that, over the long term, these challenges, combined with continuous advancement in the acquisition frequency and spatial resolution of satellite imagery (Miller et al., 2018) may render sky imagers largely obsolete in most solar forecasting applications.

5.3. Supporting concepts

In this section, abbreviations or acronyms that play supportive roles in solar forecasting are discussed. It is observed that the abbreviations in this category can be further divided into three sub-categories, namely, meteorology, statistics, and mathematics. Similarly to the earlier sections, the abbreviations for this section are tabulated in Tables 11 and 12. It is noted that many concepts in this category have a long history and rich literature. Since the descriptions of these concepts are usually brief in solar forecasting papers, due to their supportive nature only, a wealth of information is summarized here in Tables 11 and 12. Each entry is complemented with a carefully selected textbook reference, the original publication, or a reference with more focused discussion for each abbreviation, and accompanied with a reference

about solar forecasting in the “Further Reading” column.

5.3.1. Meteorology

Meteorology plays an important role in solar forecasting. Meteorology is a branch of the atmospheric sciences that focuses on weather processes and forecasting. It is the main driver of the NWP and image-based forecasting methods discussed in Section 5.2. Although weather is uncertain in general due to many degrees of freedom, some variations are deterministic or can at least be calculated with high confidence.

From a time series forecasting standpoint, what distinguishes irradiance or PV power time series from any other kind of time series is essentially the diurnal cycle due to the apparent position of the Sun. For that reason, almost every statistical and machine learning solar forecasting paper removes the diurnal cycle before building a forecasting model. There are many statistical detrending methods (see Yang (2017), for a detailed discussion), however the most common approach is by considering either the extraterrestrial irradiance or the clear-sky irradiance. The former can be calculated using a solar positioning algorithm, whereas the latter requires a clear-sky radiation model (CSM). A CSM can be either empirical or physical. Whereas the empirical models are easy to manipulate, they are often location-dependent and have lower accuracies in general (e.g., Yang et al., 2014b; Yang et al., 2012a; Janjai et al., 2011). Physical CSMs, on the other hand, usually consist of some simplified version of a physical radiative transfer model (RTM). These CSMs have varying degrees of sophistication that correlate with accuracy in general. Well-validated models with good performances include REST2 (Gueymard, 2008), McClear (Lefèvre et al., 2013), and Ineichen (Ineichen, 2008). The reader is referred to the comparison papers by Zhong and Kleissl (2015), Gueymard (2012a), and Gueymard and Ruiz-Arias (2015) for more details on RTM-based clear-sky models. The remaining frequently-used meteorological abbreviations can be classified into two groups, namely, meteorological parameters and instruments. They are summarized in Tables 9 and 10, respectively.

As shown in Fig. 5, GHI and DNI are the two most frequent abbreviations in this category, owing to the fact that they are the most influential factors affecting the power output of PV and concentrating solar power (CSP) plants, respectively. It can also be observed that many meteorological parameters shown in Table 9, such as cloud cover (CC) and cloud index (CI), are related to clouds. Cloud physics and cloud models are thought to be the most important factors contributing to solar forecast accuracies. Yet clouds are one of the most complex natural phenomena to model as they can contain all three phases of water, a combination of solids and fluids, turbulent mixing, and six orders of magnitudes in length scales. Other meteorological variables such as relative humidity (RH) and wind speed (WS) are also important to physical cloud modeling. The preferred modeling approach differs by forecast horizon.

Measurements of meteorological parameters come from two complementary sources: remote sensing and ground-based instruments. The Earth’s atmosphere is continuously sensed by a fleet of geostationary meteorological satellites, e.g., the geostationary operational environmental satellite (GOES) satellites operated by NOAA, Meteosat operated by the European Space Agency, or multi-functional transport satellite (MTSAT) operated by JMA over Asia. The Earth coverage is complete, except over high-latitude and polar regions, where such imagery is not exploitable. Using the images captured by these satellites, and an appropriate cloud-to-irradiance conversion algorithm (e.g., Qu et al., 2017; Perez et al., 2002), GHI, DNI and thus DHI estimates can be derived. Although satellite-derived irradiance provides global coverage, spatial resolution and accuracy can be limited. To that end, site-adaptation techniques—making corrections to satellite-derived irradiance using local ground-based measurements—improve the bankability of the datasets (see survey by Polo et al., 2016). When ground-measurements are considered, two types of instrument, pyranometer and pyrheliometer, are most commonly used to measure GHI and DNI, respectively. The reader is referred to the book by Vignola et al. (2012) for a tutorial on various types of pyranometers and pyrheliometers. Recently, sky images were successfully post-processed to obtain direct and diffuse irradiance estimates (Kurtz and Kleissl, 2017). Conventionally, however, sky imagers have been used only for aerosol characterization, cloud detection, and forecasting. Originally, the Total Sky Imager (TSI) produced by Yankee Environment Systems (YES) was the only commercially available system and therefore popular among researchers. Recently, however, more vendors have entered the market and researchers have preferred lower-priced security cameras for solar forecasting (Schmidt et al., 2016; Yang et al., 2014d).

5.3.2. Statistics

The word “statistics” means either the subject itself, i.e., statistical science, or summaries of data, e.g., mean or variance. In earlier sections, many statistical forecasting models have been mentioned, along with statistical error metrics. However, the word statistics here refers to those methods or concepts that fall under the subject of statistics, but do not directly relate to solar forecasting. As shown in Fig. 5, branches under statistics can be either a stand-alone concept or fall under a subcategory. In the paragraphs below, the stand-alone concepts are briefly introduced first, followed by a discussion about the three sub-categories, namely, hypothesis testing, model selection and parameter estimation. It is noted that these concepts and categories are not exhaustive; new concepts and categories can be developed when the text corpus expands. However, this preliminary study is limited to the abbreviations shown in Fig. 5 and summarized in Table 11.

A common way to introduce statistics in a graduate course is to start with probability (e.g., axioms of probability and Bayes’ theorem), random variables (e.g., various discrete and continuous random variables and their properties), expectations (e.g., variance and moment-generating function), inequalities and convergence (e.g., Chebyshev’s inequality and central limit theorem), and then inference (i.e., method of moments and maximum likelihood estimation). Although solar irradiance is a physical parameter, some statistical concepts can be very useful in describing irradiance. To that end, two statistics textbooks can be recommended (Wasserman, 2006, 2004). In these books, a structured, concise and complete introduction to statistics is provided. The learning curve to study the two books can be steep. Nevertheless, as statistics takes a major role in solar forecasting, it would be beneficial to build statistical foundation among solar forecasters. In this way, many excellent and state-of-the-art statistical methods published in the top statistics journals, such as Journal of the American Statistical Association, Journal of the Royal Statistical Society, or Annals of Statistics, can be better utilized to forecast solar irradiance. Aside from general statistics textbooks, topic-specific books, such as that of Box and Jenkins (1994), are very helpful to understand the underlying motivation and derivation of concepts such as autocorrelation function (ACF) or partial autocorrelation function (PACF).

Hypothesis testing is a major subject in statistical inference. There are many tests that can be useful in solar forecasting; according to the preliminary results obtained from the present study, the most frequently-used ones are analysis of variance (ANOVA), augmented Dickey–Fuller (ADF), and Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test. Other important tests include the Kolmogorov–Smirnov (KS) test (Massey, 1951), which tests the equality of continuous, one-dimensional probability distributions; the Diebold–Mariano (DM) test (Diebold and Mariano, 1995), which is the only test that compares the forecast accuracy of two methods; and the Wald test (Gouriéroux et al., 1982), which tests whether the estimated parameter is equal to a proposed value, e.g., whether the coefficient of a regressor is zero. Whereas the statistical testing is very powerful and provides statistical evidence for hypotheses, the subject is often abused by ignoring some important assumptions and regularizations. For instance, twenty years after the original paper (Diebold and Mariano, 1995) was proposed, the lead

Table 9
Frequently-used abbreviations in solar forecasting: support—meteorology—parameter. The first reference for most entries in this table provides a more detailed discussion on the meteorological parameter, then followed by one solar forecasting reference that typifies the usage.

Abbrv.	Long form	One-sentence description	Further reading	Confusion (C) or incorrect usage (I)
AOD	Aerosol Optical Depth	A dimensionless measure of the extinction of the solar beam by dust and haze; typical values range from 0.01 to 0.40 for very clean to hazy atmospheric conditions, but extremely high values above 1 do occur in various situations or regions	Gueymard (2012b) and Lara-Fanego et al. (2012)	–
CSI	Clear-Sky Index	The ratio between GHI and clear-sky irradiance at the surface; not to be confused with clearness index, which is the ratio between GHI and extraterrestrial/top-of-atmosphere GHI	Engerer and Mills (2014) and Voyant et al. (2017b)	California Solar Initiative (C)
CBH	Cloud Base Height	The lowest altitude of a cloud or cloud field, a critical parameter for sky-imager-based forecasting during ray tracing	Peng et al. (2015) and Chow et al. (2011)	–
CC	Cloud Cover	Aerial fractional amount of total cloudiness covering the sky, expressed in okta (by human observers), pixel value (by image-based methods), or time-equivalent fraction (by ceilometers)	Tapakis and Charalambides (2013) and Yang et al. (2014d)	Cycle Charging (C); Combined Cycle (C)
CI	Cloud Index	A component derived from satellite images, can be thought of as the reflectivity enhancement of the ground due to clouds as observed from space	Dagestad and Olseth (2007) and Arbizu-Barrena et al. (2017)	Confidence Interval (C); Clearness Index (C)
DHI	Diffuse Horizontal Irradiance	The solar irradiance scattered by molecules, aerosols, and clouds in the atmosphere and received on a horizontal surface	Gueymard (2017) and Yang et al. (2012b)	Diffuse Horizontal Irradiation (I)
DNI	Direct Normal Irradiance	Solar radiation directly reaching the Earth's surface without angular deflection	Blanc et al. (2014) and Chu et al. (2016)	Direct Normal Incidence (I); Direct Normal Irradiation (I)
DSWRF	Downward ShortWave Radiation Flux	A quantity equivalent to GHI, one of the non-native parameters from NWP. Since NWP models also compute longwave fluxes, the "shortwave" is added, while "shortwave" is often omitted in solar energy terminology.	Zhang et al. (2015c) and Perez et al. (2013)	–
GHI	Global Horizontal Irradiance	The total solar flux available from the sky dome that is incident on a horizontal surface	Vignola et al. (2012) and Perez et al. (2010)	Global Horizontal Irradiation (I)
GSI	Global Solar Irradiance	Not a commonly accepted abbreviation; the use of GHI is preferred, or global tilted irradiance (GTI) for inclined surface	–	–
PBL	Planetary Boundary Layer	The lowest part of the atmosphere, i.e., just adjacent to the Earth's surface; most optically thick clouds are driven by moisture and temperature variations in the PBL	Deardorff (1972) and Yang and Kleissl (2016)	–
RH	Relative Humidity	Amount of water vapor in the air relative to the saturation amount at the same temperature; a frequently-used exogenous variable	Lawrence (2005) and Urraca et al. (2016)	–
SZA	Solar Zenith Angle	The angle between the zenith (vertical) and the Sun; one of the two angles—the other is the azimuth angle—for solar positioning	Reda and Andreas (2004) and Huang and Davy (2016)	–
SSI	Surface Solar Irradiance	Not a commonly accepted abbreviation; the use of DSWRF or surface solar radiation downwards (SSRD) is preferred. Specific to NWP where solar irradiance at higher atmospheric levels is also of interest.	–	–
TCC	Total Cloud Cover	See CC	See further reading for CC and CI	–
WS	Wind Speed	A fundamental atmospheric quantity directly related to the movement of low clouds	Gutierrez-Corea et al. (2016)	–

Table 10
Frequently-used abbreviations in solar forecasting: support—meteorology—instrument. The first reference for most entries in this table provides a more detailed discussion on the instrument, then followed by one solar forecasting reference that typifies the usage.

Abbrv.	Long form	One-sentence description	Further reading	Confusion (C) or incorrect usage (I)
GOES	Geostationary Operational Environmental Satellite	A satellite system operated by NOAA; the first fourth-generation satellite (GOES-R) was recently launched	Polo et al. (2008) and Marquez et al. (2013)	–
MTSAT	Multi-functional Transport SATellite	A meteorological satellite operated by JMA; it has been replaced by Himawari-8 on 2015-07-07	Polo et al. (2008) and Dong et al. (2014)	Multi-purpose Transport SATellite (I)
NIP	Normal Incidence Pyrheliometer	A model of pyrheliometer—an instrument that measures DNI—produced by Eppley; it is mounted on a tracker to follow the sun	Vignola et al. (2012) and Marquez and Coimbra (2013)	–
PSP	Precision Spectral Pyranometer	A model of pyranometer—an instrument that measures primarily GHI, but that can also measure DHI if equipped with a tracking shade—produced by Eppley	Vignola et al. (2012) and Marquez and Coimbra (2013)	–
Sky imager				
CCD	Charge Coupled Device	A major technology for digital imaging by manipulating electrical charge and converting it into a digital signal, i.e., pixel values	Nakamura (2016) and Yang et al. (2014d)	–
HDR	High Dynamic Range	HDR imaging is the compositing and tone-mapping of images to extend the dynamic range beyond the native capability of the capturing device	Reinhard et al. (2010) and Urquhart et al. (2015)	–
NRBR	Normalized Red to Blue Ratio	A normalized version of RBR, $\lambda_N = (\lambda-1)/(\lambda+1)$, where $\lambda = b/r$ and b,r are blue and red channel values, $0 \leqslant b,r \leq 255$	Li et al. (2011) and Chu et al. (2016)	Normalized Red Blue Ratio (C)
RBR	Red to Blue Ratio	The ratio of red to blue light in a pixel	Pfister et al. (2003) and Ghonima et al. (2012)	Red Blue Ratio (C)
TSI	Total Sky Imager	A sky-imager system produced by YES (no longer sold)	Long et al. (2001) and Chu et al. (2013)	Total Solar Irradiance (C)
YES	Yankee Environmental Systems	A company that produces radiometry, environmental imaging, and other ground-based devices	Long et al. (2001) and Chu et al. (2013)	–

author Francis Diebold wrote another paper discussing the widespread misuse of his test across various scientific domains (Diebold, 2015). Therefore, when applying hypothesis tests in solar forecasting applications, it is recommended to not only follow solar forecasting references, but to understand the tests from a statistical point of view.

Besides hypothesis testing, the model selection problem is common to solar forecast practitioners. As factors affecting solar irradiance are numerous, the regression methods discussed in Section 5.2.2 become a popular choice for predictions, and thus model selection is needed. Model selection often involves procedures such as the design of experiments (DoE), where the effect of model inputs on its output can be studied. In other circumstances, one needs to select a model from a set of candidate models, and thus the selection criteria are of interest. Commonly used criteria in solar forecasting include Akaike information criteria (AIC), Bayesian information criteria (BIC), and structural risk minimization (SRM). AIC and BIC are useful to automatically select time series models, whereas SRM motivates SVM.

Parameter estimation is required after a forecasting model is constructed. Here, parameter refers to the part of a model that needs to be estimated from the data. As far as parametric inference is concerned, the two most widely-used parameter estimation methods are the method of moments and the maximum likelihood estimation (MLE). The former is often not optimal but easy to compute, whereas the latter benefits from many properties, such as consistency, equivariance, asymptotic normality, and optimality (Wasserman, 2004). Whereas the MLE's definition is rather simple, its computation can be difficult because some ML estimators are not analytical. For that reason, numerical methods such as the expectation–maximization (EM) algorithm proposed by Dempster et al. (1977) received over 50,000 citations. Perhaps the main drawback of parametric inference is the assumption that has to be made on the data distribution. Besides a few clear cases—an arrival process is Poisson, for example—distributions are rarely known. In practice, the least-squares (LS) estimator is widely applied to observations of any distribution. Under some conditions, its variant, the weighted LS estimator, is equivalent to the ML estimator. Nevertheless, the LS estimator has several drawbacks, such as covariance inversion problem for high-dimensional data and penalization of outliers. Fortunately, if statistical models are used in solar forecasting, the preferred parameter estimation method is often known. Many statistical software packages also offer readily available toolkits. Lastly, cross validation (CV) provides an accuracy-based estimation, and is often used to tune certain parameters in a model, e.g., the penalty strength in a LASSO model.

5.3.3. Mathematics

Various concepts and tools in the field of mathematics have been applied to solar forecasting, with numerical optimization[16] being the most applied tool. Optimization is an important tool in decision science and in the analysis of physical systems (Nocedal and Wright, 1999). There are many ways to classify an optimization problem, e.g., continuous or discrete, constrained or unconstrained, global or local, and stochastic or deterministic optimization. Most abbreviations related to numerical optimization are triggered by specific needs of the underlying solar forecast algorithm. For example, quadratic programming (QP) and Karush–Kuhn–Tucker (KKT) are mentioned in conjunction with SVM; the minimum cross-entropy (MCE) method is mentioned when cloud identification through sky images is needed; similarly, the Levenberg–Marquardt (LM) algorithm is mentioned when a back-propagation neural network is used. Optimization abbreviations frequently appear when they are used to solve for some parameters in a forecasting model. For instance, the Broyden–-Fletcher–Goldfarb–Shanno (BFGS) method and simulated annealing

[16] Note that numerical optimization herein refers to the classic methods such as the Newton–Raphson method, but not to heuristics in artificial intelligence.

Table 11
Frequently-used abbreviations in solar forecasting: support—statistics. The first reference for most entries in this table comes from statistics (mostly text books), then followed by one solar forecasting reference that typifies the application.

Abbrv.	Long form	One-sentence description	Further reading	Confusion (C) or incorrect usage (I)
ACF	AutoCorrelation Function	A function (x is time lag, y is correlation) or a set of statistics (e.g., first- and second-order autocorrelations) to describe the correlations between time series points separated by various time lags	Box and Jenkins (1994) and Yang et al. (2017a)	AutoCorrelation Coefficient (I); Auto Correlation Function (I)
CDF	Cumulative Distribution Function	A function that maps real values (random variables) to the range [0,1],$F_X(x) = \mathbb{P}(X \leqslant x)$	Wasserman (2004) and Espinar et al. (2009)	Cumulative Density Function (I)
GP	Gaussian Process	A model for data in a continuous space and/or time domain, in which every point is normally distributed, and any finite collection of these variables has a multivariate normal distribution	Rasmussen and Williams (2006) and Lauret et al. (2015)	Genetic Programming (C)
GAM	Generalized Additive Model	A generalization of linear models, in which each linear term in a GLM is replaced by some unknown smooth function	Hastie and Tibshirani (1990) and Paulescu et al. (2017)	–
KDE	Kernel Density Estimation	A nonparametric probability density estimation method, can be thought of as a smooth version of histogram	Wasserman (2006) and van der Meer et al. (2017)	–
MOS	Model Output Statistic	An MLR method in which the predictand is related statistically to one or more predictors, often used to correct results from NWP	Yang and Kleissl (2016) and Mathiesen and Kleissl (2011)	Multiple Output Statistic (I)
MDS	MultiDimensional Scaling	A space transformation technique, where the distances between points in the low dimensional space match the original dissimilarities in data	Cox and Cox (2000) and Yang et al. (2013b)	–
PACF	Partial AutoCorrelation Function	A function or a set of statistics to describe the autocorrelations between time series points separated by various lags, conditioning on the observations between these pairs of points	Box and Jenkins (1994) and Bouzerdoum et al. (2013)	Partial AutoCorrelation Coefficient (I)
PCA	Principal Component Analysis	An orthogonal transformation that converts a set of observations to a set of linearly uncorrelated variables	Jolliffe (1986) and Yang et al. (2017a)	Principal Components Analysis (I)
PDF	Probability Density Function	Describes the distribution of continuous random variables; $f_X(x) = F'_X(x)$ at all points x where CDF is differentiable	Wasserman (2004) and van der Meer et al. (2017)	Probability Distribution Function (I)
Hypothesis testing				
ANOVA	ANalysis Of VAriance	A statistical technique for analyzing measurements depending on several kinds of effects operating simultaneously, to decide which kinds of effects are important and to estimate the effects	Sahai and Ojeda (2004) and Zhang et al. (2015a)	–
ADF	Augmented Dickey Fuller	An augmented autoregressive unit root test for general ARMA with unknown orders	Dickey (2011) and Raza et al. (2016)	–
KPSS	Kwiatkowski Phillips Schmidt Shin	The most commonly-used stationarity test for time series	Kwiatkowski et al. (1992) and Dong et al. (2013)	–
Model selection				
AIC	Akaike Information Criterion	A penalized model (e.g., time series models such as ARIMA, ETS families) selection method based on in-sample fit; its computation requires the likelihood function and the number of parameters	Hyndman et al. (2008) and Yang et al. (2012b)	–
BIC	Bayesian Information Criterion	Same as AIC, except that the penalty is now also a function of sample size	Hyndman et al. (2008) and Li et al. (2014)	–
DoE	Design of Experiment	A method to explore the relationship between factors affecting a process and the output of that process	Fisher (1937) and Zhang et al. (2015a)	Department of Energy (C)
GDF	Generalized Degrees of Freedom	Defined as the sum of sensitivities of fitted values with respect to the observed response values; a theory that allows complex modeling procedures to be analyzed in the same way as linear models	Ye (1998) and Urraca et al. (2016)	–
SRM	Structural Risk Minimization	An inductive principle for model selection used for learning from finite training data; SVM uses the SRM principle	Zhang (2011) and Wang et al. (2015)	–
Parameter estimation				
CV	Cross Validation	A method in statistics and machine learning to estimate parameters based on partitioned datasets; it is also used for model selection and results validation	Wasserman (2006) and Chu et al. (2015b)	Computer Vision (C); Coefficient of Variation (C)
EM	Expectation Maximization	An iterative method to compute maximum likelihood estimates, often used when it is difficult to perform MLE analytically	Wasserman (2004) and Soubdhan et al. (2016)	–
GCV	Generalized Cross Validation	A rotation-invariant version of ordinary cross validation; commonly used in nonparametric regression	Golub et al. (1979) and Li et al. (2016c)	–
LS	Least Squares	Methods to estimate parameters in a regression model; its variants include OLS, weighted least squares and generalized least squares	Kariya and Kurata (2004) and Yang et al. (2017b)	–
MLE	Maximum Likelihood Estimation	A method to estimate a parameter by maximizing the likelihood function defined as $\mathscr{L}_n(\theta)\prod_{i=1}^{n} f(X_i;\theta)$, where n is sample size; f is the PDF; and θ is the parameter to be estimated	Wasserman (2004) and Bouzerdoum et al. (2013)	Maximum Likelihood Estimates (I)

Table 12
Frequently-used abbreviations in solar forecasting: support—mathematics. The first reference for each entry in this table comes from mathematics (mostly text books), then followed by one solar forecasting reference that typifies the application.

Abbrv.	Long form	One-sentence description	Further reading	Confusion (C) or incorrect usage (I)
MI	Mutual Information	A measure of mutual dependence between two variables; it is used to decide the maximum of lagged inputs to consider in solar forecasting	Shannon (1948) and Voyant et al. (2017a)	–
PFA	Probabilistic Finite Automata	A non-deterministic mathematical model for information processing that returns responses based on some probabilities	Salomaa (1969) and Mora-López et al. (2011)	–
Optimization				
BFGS	Broyden Fletcher Goldfarb Shanno	A quasi-Newton method for solving unconstrained nonlinear optimization problems	Fletcher (2013) and Sperati et al. (2016)	–
KKT	Karush Kuhn Tucker	A set of first-order necessary conditions to assure the optimality of a solution in a nonlinear programming problem; in the solar forecasting literature, it is used to find the bias parameter in an SVM	Bhatti (2000) and Li et al. (2016a)	–
LM	Levenberg Marquardt	An algorithm to solve nonlinear least-squares optimization problems; it commonly appears in the solar forecasting literature as a backpropagation neural network training method	Nocedal and Wright (1999) and Ahmad et al. (2015)	Lagrange Multiplier (C); Linear Model (C)
MCE	Minimum Cross-Entropy	An optimization method for combinatorial optimization problems and rare-event probability estimation; its primary application in solar forecasting is cloud identification for sky images	Li and Lee (1993) and Li et al. (2016b)	–
MILP	Mixed Integer Linear Programming	MIP with a linear objective function	See MIP	–
MIP	Mixed Integer Programming	An optimization problem over some integer-valued decision variables and some real-valued variables; usually not directly involved in forecasting, but used in the subsequent operations, including economic dispatch and unit commitment	Wolsey (2007) and Lujano-Rojas et al. (2016)	–
QP	Quadratic Programming	An optimization problem that involves a quadratic objective function subject to bounds, linear equality, and inequality constraints; in the solar forecasting literature, it is often used to find the coefficients of kernels in an SVM	Lee et al. (2005) and Lauret et al. (2015)	–
SA	Simulated Annealing	A general-purpose, serial algorithm for finding the global minimum of a continuous function	Kirkpatrick et al. (1983) and Akarslan and Hocaoğlu (2016)	Sensitivity Analysis (C)
Transform				
CWT	Continuous Wavelet Transform	An integral transform similar to Fourier transform, but with a kernel that is a function of both scale (a, the scaling parameter) and location (b, the translation parameter)	Debnath and Shah (2015) and Monjoly et al. (2017)	–
DWT	Discrete Wavelet Transform	The discrete version of CWT that assumes a and b take only integral values, usually $a = 2^m$ and $b = n2^m$, where m and n are integers	Debnath and Shah (2015) and Sharma et al. (2016)	Discrete Wavelet Transformation (I)
FFT	Fast Fourier Transform	A numerical algorithm that allows fast computation of (discrete) Fourier transform; primarily used for variability studies prior to forecasting	Debnath and Shah (2015) and Dong et al. (2013)	–
PIT	Probability Integral Transform	A theorem stating that if a random variable X has a continuous distribution function $F(x)$, then the random variable $U = F(X)$ has a uniform distribution over the interval $(0,1)$; can be used as a metric for probabilistic forecast	Quesenberry (2006) and Verzijlbergh et al. (2015)	–
WT	Wavelet Transform	The wavelet transform can be divided into two categories: CWT and DWT	See CWT and DWT	Wind Turbine (C)

(SA) are two commonly used methods. The choice of optimization algorithms may largely depend on their availability in software packages; for instance, both BFGS and SA are implemented in the most fundamental optimization function **optim** in **R**. Lastly, some solar forecast papers also cover numerical optimization of grid operations using the forecasts; for example, the mixed integer programming (MIP) is often used for economic dispatch and unit commitment.

In a similar manner as optimization, mathematical transform—another very broad term—can be classified in many ways. In solar forecasting, Fourier transform (FT) and wavelet transform (WT), or their variants, are the two most used transforms. Both can be classified as integral transforms: $F(p) = \int_a^b K(p,x)f(x)dx$, where $f(x)$ is the function under transformation and $K(p,x)$ is the kernel of the transformation. The main motivation for a transformation is that the determination or manipulation of $F(p)$ is often more convenient than that of $f(x)$ (Miles, 1971). Furthermore, the transformation also reveals characteristics and features of the time series that are otherwise not observable in its original time domain. Mathematical transforms have been widely used to study the variability in irradiance time series (e.g., Lave et al., 2013). Since the variability can be modeled through transformations, Fourier and wavelet decompositions can be used to remove the time series trend at some specific frequency (Dong et al., 2013), and to create sub-series that are easier to forecast (Zhu et al., 2017). Extended applications of the wavelet transform include wavelet neural network, which uses wavelets as activation functions (Sharma et al., 2016; Mellit et al., 2006), and wavelet-coupled SVM, which decomposes the input signal before applying SVM to produce forecasts (Deo et al., 2016).

5.4. System, software and data

With the rapid uptake of solar forecasting, many countries and organizations have developed forecasting systems to suit various research and operational needs. Some of these forecasting systems are designed for general grid integration purposes, e.g., the Australian Solar Energy Forecasting System (ASEFS), whereas others are designed for very specific tasks, e.g., smart adaptive cloud identification system (SACI) and integrated solar forecasting platform (ISFP), see Table 13. Beside systems, software and data have also been developed and collected. Most forecasting systems and datasets being proprietary, it is useful to know more about the freely available ones. Aside from discussing the frequently-used abbreviations, this section also reviews some online available datasets that can be used to validate forecasting studies, as well as some common research and production languages to implement forecasting algorithms.

5.4.1. Online available datasets

Big data, characterized by 5 Vs[17] (Ishwarappa and Anuradha, 2015), is one of the hottest research topics in today's world. In a recent publication (Haupt and Kosović, 2017), these characteristics are associated with the nature of solar forecasting data. A parallel definition for big data is described by the HACE theorem[18] (Wu et al., 2014). Datasets for solar forecasting are also well-aligned with the HACE theorem.[19] For such reasons, an increasing number of publications use data from more than one source to perform solar forecasting; this forms a large body of literature on hybrid methodologies. It is therefore important to become aware of various online available datasets that may be complementary to the data at hand. Another reason for using the available datasets online is that the self-collected data are often incomplete, the sample size is too small, or poorly controlled, which is a common reason that leads to rejection of a manuscript by a journal (e.g., see the editorial by Gueymard et al., 2009).

Searching for online available datasets can be time consuming. Fortunately, several reviews have been conducted earlier (e.g., Sengupta et al., 2017; Paulescu et al., 2013; Gueymard and Myers, 2008). In these reviews, the popular databases such as the Baseline Surface Radiation Network (BSRN), European Solar Radiation Atlas (ESRA), and National Solar Radiation Database (NSRDB) are described in detail.

Besides the databases, another emerging source for solar forecasting data is the supplementary materials to the publications. In many statistics journals such as Journal of Computational and Graphical Statistics or Technometrics, it is customary to publish data and code as supplementary materials, especially when the computation and implementation details are difficult to replicate. Such action not only improves the readability of the papers, but also stimulates further research on the same topic using the same data. With the obvious benefits of data sharing, many journals from other fields also have similar policies. Some journals even made it compulsory to submit data and code, e.g., Marketing Science, a top journal in the area of operations research and management science. The primary objective of such policy is to ensure the replicability of the published papers (see Desai (2013), for detailed discussion on why and how Marketing Science makes this policy possible). In recent years, several authors have also started to submit solar forecasting related data and code to Solar Energy, for example, Dazhi Yang (Yang et al., 2017a,b,c, 2015d, 2014c; Yang, 2017, 2016) or Antonio Lorenzo (Lorenzo et al., 2017, 2015). Other authors also take the approach to the next level—Oscar Perpiñán and Marcelo Almeida in particular—by publishing general-purpose open-source software packages, based on their solar forecasting publications (e.g., Almeida et al., 2017, 2015). This policy appears beneficial at large and is encouraged here. In this way, more forecasting-related data can be made available, and thus can facilitate international research competition and collaboration, which is important for scientific progress.

Besides the above-mentioned ways to find available datasets online, the most effective method is believed to be text mining. By searching keywords such as "http", "ftp", "freely available", "freely accessible at", "data and code"[20] in a large collection of solar forecasting texts, sentences containing the URLs of online databases may be discovered. A demonstration of such searches will appear in a subsequent contribution.

5.4.2. Programming languages, software and databases

IEEE Spectrum has conducted interactive ranking of programming languages based on 12 metrics from 10 sources on a yearly basis since 2014. Currently (2017), the top 10 languages are **Python**, **C**, **Java**, **C++**, **C#**, **R**, **Javascript**, **PHP**, **Go** and **Swift**, in that order.[21] While some new languages enter the top 10 list, other languages that were once on the list, such as **Objective-C** and **Matlab**, have dropped to lower positions. Among the top 10 languages, two languages in particular can be recommended for solar forecasting research, namely, **Python** and **R**.

Generally speaking, solar forecasting algorithms can be written in any language. Here, **Python** and **R** are recommended because they can carry out many tasks, such as sky imagery processing, time series analysis, machine learning, or plotting. The reasons for this recommendation include their open-source nature, high compatibility

[17] Almost all descriptions at least involve three Vs: volume, velocity, and variety. The other two Vs are veracity and value; sometimes, veracity is replaced by variability.

[18] Big data starts with large-volume, **h**eterogeneous, **a**utonomous sources with distributed and decentralized control, and seeks to explore **c**omplex and **e**volving relationships among data.

[19] (Heterogeneous) Various solar radiation databases include data of diverse dimensionality. They are collected under different schemata and protocols. (Autonomous sources) Data used for solar forecasting comes from various autonomous channels, including satellite, sky cameras, and other ground-based sensors. (Complex and Evolving) The relationships among various types of data are complex and complementary. The schemata, protocols and devices for data gathering are constantly evolving.

[20] Only searching for the word "data" is not useful. However, when "data" and "code" are searched together, the chance of finding an available dataset is higher.

[21] https://spectrum.ieee.org/computing/software/the-2017-top-programming-languages.

Table 13
Frequently-used abbreviations in solar forecasting: systems, software and databases.

Abbrv.	Long form	One-sentence description	Further reading	Confusion (C) or incorrect usage (I)
AFSOL	Aerosol-based Forecasts of Solar Irradiance for Energy Application	A system that produces GHI and DNI forecasts, covering Europe and the Mediterranean region, particularly accurate for clear-sky irradiance	Breitkreuz et al. (2009) and Law et al. (2014)	–
AERONET	AErosol RObotic NETwork	A network of monitoring stations that extract properties of the atmospheric column such as aerosol optical depth	Urraca et al. (2016) and Breitkreuz et al. (2007)	–
ASEFS	Australian Solar Energy Forecasting System	A forecast system developed for Australian Energy Market Operator to enhance integration of solar energy generation at all scales into the Australian national grid	Grantham et al. (2016)	–
BSRN	Baseline Surface Radiation Network	A global network of meteorological stations that make high-quality shortwave and longwave radiation observations; these data are often used for model validation	Boilley et al. (2016) and Lara-Fanego et al. (2012)	–
CIMIS	California Irrigation Management Information System	A network of agricultural weather stations in California that measure Global Horizontal Irradiance with silicon pyranometers	Yang and Kleissl (2016) and Zagouras et al. (2015)	–
ESRA	European Solar Radiation Atlas	The Atlas contains both a database and software packages, to access the solar potential in Europe; often quoted for the clear-sky model therein	Scharmer et al. (2000) and Verzijlbergh et al. (2015)	–
GIS	Geographic Information System	A multidisciplinary technology for the collection, storage, manipulation, analysis and visualization of spatial information; mainly used during solar resource assessment	Karakaya (2016) and Ramirez-Rosado et al. (2011)	Geospatial Information System (I); Geographical Information System (I)
ISFP	Integrated Solar Forecasting Platform	A short-term GHI and DNI forecasting platform developed by UCSD that takes cloud information as exogenous inputs to ANN	Chu et al. (2015b)	–
LES	Large Eddy Simulation	A simulation model originally developed for atmospheric flow prediction; it is now applied to other turbulent flow problems in engineering	Yang (2015) and Verzijlbergh et al. (2015)	Linear Exponential Smoothing (C)
NSRDB	National Solar Radiation DataBase	A database containing meteorological and solar irradiance data for the United States and some other parts of the American continent	Nonnenmacher et al. (2016) and Jiang et al. (2015)	–
SACI	Smart Adaptive Cloud Identification system	A system developed by UCSD to conduct cloud detection via fish-eye cameras; it integrates fixed threshold methods, MCE, and clear-sky library	Li et al. (2016b) and Chu et al. (2015b)	–
SUNY	State University of New York GOES satellite based solar model	The most established model for satellite-derived irradiance; it provides half-hourly data for the United States for 1998 to present and for parts of South Asia	Perez et al. (2002) and Law et al. (2014)	–
SCADA	Supervisory Control and Data Acquisition	A high-level process supervisory management system that consists of components including supervisory computers, remote terminal units, communication devices, programmable logic controller, and user interface	Sepasi et al. (2017) and Chen et al. (2011)	–
SAM	System Advisor Model	A software package developed by NREL for renewable energy system performance simulation and financial viability studies	Law et al. (2016a) and Law et al. (2016b)	–
TMY	Typical Meteorological Year	A dataset that typifies the climatic and weather conditions of a location, often used for simulating building energy use and solar power production	Wilcox and Marion (2008) and Yang et al. (2015c)	–

* AERONET, BSRN, CIMIS, and ESRA could also be listed in Table 14.

with hardware and other software, moderate learning curve, rich online support (forums, sample code and packages written by peers), no compilation required, ability to handle complex tasks, fast realization of algorithms, and strong visualization capabilities.[22] In recent years, an increasing number of models and analyses in various energy journals have used **Python** and **R**; the latter is often accompanied with the eye-catching signature plots with grey background and gridlines produced using the **ggplot2** package. Furthermore, there are many **Python** and **R** packages written specifically for solar resources and forecasting applications, e.g., **Solpy** and **PVLIB** in **Python**, or **meteoForecast** and **solaR** in **R**. These packages contain standard implementations of many popular radiation models, as well as interfaces to various data sources, which can be very helpful.

Besides programming languages, there are other free or commercially-packaged software tools that can be downloaded and installed. Example abbreviations, such as geographic information system (GIS) software (e.g., **ArcGIS** or **GRASS GIS**), and **System Advisor Model** (SAM) are listed in Table 13. In general, these software tools play a supporting role in solar forecasting, namely for resource assessment and irradiance-to-power conversion.

5.5. Organizations

It is found that abbreviations for organization names often appear in the full texts. The purpose of having these abbreviations listed here is similar to that of Section 4.4, namely, to facilitate research collaborations and identify potential data sources. Some organization abbreviations shown in Fig. 5 are explained in Table 14. The remaining ones are omitted, since abbreviations such as Department of Energy (DOE) or National Science Foundation (NSF) are mostly mentioned in the acknowledgement section of the papers. All contact information listed in Table 14 is obtained based on publicly available search.

6. Emerging technology in solar forecasting

Following the voting procedure described in Section 3.2.3, the ranking results for the top six publications (mean rank $\geqslant 3$) are shown in Table 15. There are several honorable mentions, including Gulin et al. (2017), Arbizu-Barrena et al. (2017), Massidda and Marrocu (2017) and Pierro et al. (2016), which have a mean rank very close to 3, or have obtained the highest rank from one of the voters. In consideration of the potential interest in analyzing these honorable mentions, as well as other publications, the code used in this section is provided as supplementary material.

For each of the winning publications, the PDF version is downloaded from ScienceDirect. The text preprocessing sequence on these PDF files is described below:

1. Read in PDF files using **Poppler**.[23] Since **Poppler** reads PDF page-by-page and stores all pages as a list, these are concatenated at this stage.
2. Translate unicode Latin ligatures.
3. Remove all text in the reference section by locating the last appearance of “References” in the text, and removing the text thereafter.
4. Remove all text in the acknowledgement section by locating the last appearance of “Acknowledgement” in the text, and removing the text thereafter.
5. Remove newline characters that break words during typesetting, e.g., restoring the word “forecast” from the character string `for-\necast`.
6. Split the text into lines via the newline characters, so that lines with a single word can be removed—this can effectively remove tables from the text.
7. Find and replace abbreviations with their long forms.
8. Preserve words connected by hyphen(s), by changing the hyphen(s) in those words to underscore.
9. Remove all non-alphabetic—does not include underscore—characters (alternatively, as in Section 6.3, the full stops are preserved, so that the documents can be broken into sentences).
10. Remove words with less than four characters using **R** function **gsub** and regular expressions.
11. Convert upper case to lower case.
12. Remove stop words defined in the **SMART**[24] library.
13. Remove extra whitespaces originated from the previous preprocessing steps.
14. Convert all plural forms to singular forms using javascript library **pluralize.js**.[25]

6.1. Short description for emerging technologies

Before performing text mining using the preprocessed text files, a short description of each identified emerging technology is given below.

6.1.1. Advection with ground sensors (Inage, 2017)

Sensor network-based methods are believed to have great potential in solar forecasting applications, especially for intra-hour and hourly forecasts, due to the low resolution and accuracy of most satellite-derived irradiance products available today. Besides the MLR method discussed in Section 5.2.2, other methods can also be used to exploit the spatio-temporal properties of the irradiance or PV output data collected by a sensor network. In the field of spatio-temporal statistics, spatio-temporal kriging, which uses a covariance function to describe the spatio-temporal process, is considered to be a major method for spatio-temporal prediction (Cressie and Wikle, 2011). It has been applied several times in solar forecasting (e.g., Jamaly and Kleissl, 2017; Perez et al., 2016; Aryaputera et al., 2015b; Yang et al., 2014a, 2013b). According to Cressie and Wikle (2011), however, another major method, namely, prediction using the stochastic partial differential equation (PDE), has not been well applied to solar sensor network data, at least until the recent publication by Inage (2017).

The PDE considered by Inage (2017) is the advection equation, i.e., an advection–diffusion equation without the diffusion portion of the process. In the first step, irradiance or PV output measurements from an irregular monitoring network are first interpolated onto a regular grid. The interpolation models the geostatistical process as a lattice process. As a result, a discrete time, discrete space approximation of the PDE can be used. Aside from the PDE approach, another notable contribution of the paper is the expression of an equivalence between the PDEs and machine learning. More specifically, the one-step-ahead and multi-step-ahead predictions using the advection equation are expressed as simple perceptrons and a deep-learning structure, respectively.

6.1.2. Standardizing forecast evaluation (Vallance et al., 2017)

Currently in solar forecasting, there is no standard way of assessing forecast accuracies. This is evident from the abundant error metrics discussed in Section 5.1. Solar forecast accuracy depends on a variety of factors including, but not limited to, forecast horizon, geographical location, temporal resolution, amount of data, and appropriateness of the data. However, the dominant factor is the forecasting method. Since

[22] These modern languages might be slow for computer-intensive meteorological applications in general and NWP-type forecasting in particular. For those codes that typically run on supercomputers or clusters, such as WRF, **FORTRAN** is still the language of choice because of its speed, legacy routines, and deep roots in the scientific community.

[23] https://poppler.freedesktop.org/.

[24] http://jmlr.csail.mit.edu/papers/volume5/lewis04a/a11-smart-stop-list/english.stop.

[25] https://github.com/blakeembrey/pluralize.

Table 14
Frequently-used abbreviations in solar forecasting: organizations.

Abbrv.	Long form	Relevance to solar forecasting	Contact person or data links
AEMO	Australian Energy Market Operator	The website allows free download of half-hourly aggregated electricity price and demand data of five states in Australia, dated back to 1998	–
BOM	Bureau of Meteorology	A rich database of meteorological data (see Deo and Şahin (2017), Law et al. (2016b), and Li et al. (2016a), for details); some data can be used together with AEMO data	Modeled: http://www.bom.gov.au/climate/how/newproducts/IDCJAD0111.shtml; Measured: http://www.bom.gov.au/climate/data/oneminsolar/about-IDCJAC0022.shtml
ECMWF	European Centre for Medium Range Weather Forecast	A research institute and an operational service that provide free or subscription-based NWP forecast datasets; this is a major source of data	http://apps.ecmwf.int/datasets/
EUMETSAT	European Organisation for the Exploitation of Meteorological Satellite	A wide range of meteorological and climate monitoring data and products collected by Meteosat and other satellites; the satellite application facility on climate monitoring (CM SAF) under EUMETSAT is a major source of satellite-derived data about clouds and solar radiation	https://www.eumetsat.int/website/home/Data/index.html; CM SAT: http://www.cmsaf.eu/EN/Home/home_node.html
JMA	Japan Meteorological Agency	Meteorological satellite data can be purchased	Japan Meteorological Business Support Center jmbsc@jmbsc.or.jp
JRC	Joint Research Center	The European Commission's science and knowledge service; its photovoltaic geographical information system is a source of free satellite-derived radiation data	http://re.jrc.ec.europa.eu/pvg_tools/en/tools.html
NASA	National Aeronautics and Space Administration	An agency of the United States federal government responsible for the civilian space program, as well as aeronautics and aerospace research; it provides free global forecasts based on the GEOS-5 research model	https://gmao.gsfc.nasa.gov/forecasts/
NCAR	National Center for Atmospheric Research	NCAR developed one of the most popular mesoscale models, namely, WRF, and NCAR's research data archive contains weather and climate data that can be searched and explored in a variety of ways	https://rda.ucar.edu/
NCEP	National Centers for Environmental Prediction	An office under NOAA; it runs several major NWP models including GFS and NAM	GFS: http://www.nco.ncep.noaa.gov/pmb/products/gfs/; NAM: http://www.nco.ncep.noaa.gov/pmb/products/nam/
NOAA	National Oceanic and Atmospheric Administration	An American scientific agency within the United States Department of Commerce; National Weather Service is a major line office—among a total of six major line offices—that deals with solar forecasting	https://rapidrefresh.noaa.gov/hrrr/
NREL	National Renewable Energy Laboratory	A world-leading research lab in renewable energy; it provides various solar resource data through, for example, NSRDB and Measurement and Instrumentation Data Center	https://www.nrel.gov/rredc/solar_data.html
NEDO	New Energy and Industrial Development Organization	It provides TMY datasets at 837 sites in Japan (METPV-11) in both web and download versions	http://www.nedo.go.jp/library/nissharyou.html
SMUD	Sacramento Municipal Utility District	Californian utility that owns a network of 74 irradiance sensors, as seen in Bartholomy et al. (2014)	–
SERIS	Solar Energy Research Institute of Singapore	A research institute that developed a network of ~30 irradiance sensors distributed across the island of Singapore, as described in Nobre et al. (2016)	T. Reindl thomas.reindl@nus.edu.sg
SIAR	Spanish Agency for Irrigation in Agriculture	An agency that provides daily, weekly or monthly meteorological data across Spain, as used in Urraca et al. (2016)	http://eportal.mapama.gob.es/websiar/Inicio.aspx (In Spanish)
SoDa	Solar radiation Data	A repository of various solar radiation data products and services	http://www.soda-pro.com/
TEP	Tucson Electric Power	An electric utility company that funds research projects that led to various publication by University of Arizona (e.g., Lorenzo et al., 2015; Lonij et al., 2013); it also supplied the data of a 25 MWp PV system to Zhang et al. (2015b)	–
UCSD	University of California San Diego	Harbors two major research groups covering solar forecasting, resourcing and integration	Prof. J. Kleissl jkleissl@ucsd.edu and Prof. C. F. M. Coimbra ccoimbra@ucsd.edu
UQC	University of Queensland Centre	A building that has a 433.44 kWp PV system on its rooftop; data has been used by Rana et al. (2016), Rana et al. (2015)	Prof. T. Saha: saha@itee.uq.edu.au

Table 15
Ranking results for emerging technologies in solar forecasting. Results with a mean rank <3 are omitted.

Candidate	Title	Voter 1	Voter 2	Voter 3	Voter 4	Voter 5	Mean rank
Yang et al. (2017b)	Reconciling solar forecasts: Geographical hierarchy	9	10	0	0	9	5.6
Vallance et al. (2017)	Towards a standardized procedure to assess solar forecast accuracy: A new ramp and time alignment metric	1	4	0	7	10	4.4
Inage (2017)	Development of an advection model for solar forecasting based on ground data first report: Development and verification of a fundamental model	3	7	0	3	4	3.4
Sanfilippo et al. (2016)	An adaptive multi-modeling approach to solar nowcasting	0	0	8	9	0	3.4
Kuhn et al. (2017)	Shadow camera system for the generation of solar irradiance maps	10	0	0	1	5	3.2
Killinger et al. (2017)	QCPV: A quality control algorithm for distributed photovoltaic array power output	0	9	0	0	6	3.0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

it is impossible to benchmark a new method to all existing methods, the lack of standardization in forecast evaluation poses a serious problem during the review of a manuscript. In other words, authors often choose error metrics and benchmarking models that are advantageous to them, and thus it is difficult for the reviewers and readers to compare results in an objective manner. In this regard, several previous attempts have been made by proposing metrics that are less sensitive to data (e.g., Zhang et al., 2015a; Beyer et al., 2009). In particular, the forecast skill proposed by Marquez and Coimbra (2012) has become popular,[26] and is strongly recommended here.

Recently, Vallance et al. (2017) provided a thorough discussion on the limitations of the conventional error metrics, such as RMSE, with concrete scenarios. These authors underlined that the ability to forecast two pieces of information, namely, lag and ramp, is not well characterized by conventional error metrics. In other words, two methods producing the same RMSE may differ largely in terms of lag and ramp forecasting. To that end, two new metrics, namely, temporal distortion mix (TDM)[27] and ramp score, are proposed in that paper to address the abilities to forecast lag and ramp explicitly. TDM is based on the temporal distortion index (Frías-Paredes et al., 2016), which is based itself on a dynamic time warping. A method with high-lag tendency, e.g., persistence, results in a high TDM. On the other hand, the ramp score is developed based on the swinging-door algorithm; it measures how well the ramp events in solar data can be forecast. Both metrics proposed by Vallance et al. (2017) are important criteria to assess solar forecasts. It is also worth mentioning that the *radar chart*—the "spider web" often used to compare skills of football players—used in the paper provides a great visual assistance in comparing forecasting methods.

6.1.3. Hierarchical forecasting (Yang et al., 2017b)

Almost all publications on solar forecasting focus on forecasting irradiance or PV power for a particular time horizon over a particular geographical area. However, the interaction among these forecasts at different scales is rarely being studied. In fact, different players in a renewable energy supply chain are often exposed to data with different granularities. For instance, independent system operators (ISOs) have access to net generation data over a district, and PV system owners usually have string-level power production data. Utilizing information across a supply chain and revising forecasts made at various individual levels have shown economic and operational benefits in producing forecasts for fast-moving consumer goods (Yang et al., 2016, 2015a,b), tourism (Hyndman et al., 2011; Athanasopoulos et al., 2009) and other operations management applications. It is important to consider information sharing in solar forecasting as well. Since solar energy generation naturally forms a hierarchy that consists of levels such as transmission zones, distribution nodes, PV plants, subsystems and inverters, hierarchical reconciliation is believed to bring benefits, and thus new practices and policies, to solar forecasting.

In a recent paper by Yang et al. (2017b), an optimal—in terms of forecast variance minimization—forecast reconciliation technique was used to generate revised forecasts across two geographical hierarchies, a transmission zone level hierarchy and a PV plant level hierarchy.[28] In the transmission zone level hierarchy, it is shown that the revised forecasts outperforms NWP forecasts produced by a commercial provider by significant margins on all levels. In the PV plant level hierarchy, the reconciled forecasts are shown to be better than LASSO forecasts produced by Yang et al. (2015d). As reconciliation produces much improved forecasts over the state-of-the-art forecasting methods by using information sharing exclusively, i.e., no exogenous inputs, the results are encouraging. Hierarchical forecasting should be further studied so that it can be used in operational forecasting.

6.1.4. Forecasting with multi-modeling (Sanfilippo et al., 2016)

It is known, *a priori*, that no single method can *consistently* produce better forecasts than other methods. This is best understood in the stock market. In the field of economics and finance, the assumption of a single market mechanism can be relaxed in favor of a regime-switching model, i.e., the model coefficients are different in each regime to account for multiple mechanisms (Gray, 1996). In solar forecasting, training time series models with moving windows is commonly used (e.g., Reikard et al., 2017), so that the parameters can be, or at least believed to be, most appropriate for the next forecast. In a more general sense, such practice is known as multi-modeling—a definition used by Sanfilippo et al. (2016). Multi-modeling is exemplified by the recent advance made by Haupt and Kosović (2017), where outputs from seven NWP models including GFS, NAM, RAP, GEM, WRF-Solar and two versions of high resolution rapid refresh (HRRR) models run in parallel by National Centers for Environmental Prediction (NCEP) and Earth System Research Laboratory (ESRL), are weighted based on recent performance to form a final forecast.

Similarly to Haupt and Kosović (2017), Sanfilippo et al. (2016) considered an adaptive multi-modeling approach for up to 15-min-ahead solar nowcasting. A total of four statistics or machine learning component models are considered in that work, including two AR models, an SVM, and a persistence model. A supervised classification approach is used to identify the best performing component model under a certain specification. The inputs to the classifier include the time information (month, day, hour, minute), a time series of the clearness index (a zenith angle-independent version proposed by Perez et al., 1990), and the forecast horizon, while the output of the classifier is the model with the smallest rRMSE. It is found that the multi-modeling approach has an overall 19% forecast skill over the best single model, and 45% over persistence.[29]

6.1.5. Quality control for PV power output (Killinger et al., 2017)

Although affordable ground-based irradiance sensors exist on the market, it is still not practical to install sensors at every PV system location for monitoring and forecasting purposes (Yang, 2017). Therefore, a workaround is to treat PV system as sensors and directly forecast PV power output (e.g., Lonij et al., 2013). To ensure that the PV output data can be used in a variety of algorithms designed for irradiance forecasting, metadata including system capacity, geographical location and orientation are needed.[30] However, these metadata often contain errors and uncertainties. Moreover, the power output data also need to undergo rigorous quality control before they can be used in forecasting. To improve the situation, Killinger et al. (2017) presented a method to validate and complete the system metadata, as well as an algorithm (named QCPV), for power output data quality control.

In the first step, metadata (azimuth, tilt and an overall measure of degradation) of a PV system is found via a clever nonlinear regression[31]

[26] There are ways to abuse this metric. For instance, it can be evaluated based on a raw persistence model—persistence without the clear-sky adjustment—that leads to a seemingly better result. It is advised to rigorously specify the persistence calculation.

[27] Important note: TDM is found to be sensitive to the length of error time series, likely due to the mechanisms of dynamic time warping (DTW). Setting a window size for DTW is a standard remedy (Sakoe and Chiba, 1978). However, in such cases, the values of TDM can be manipulated by the window size. On this point, an improved version of TDM should be sought. At this stage, the authors advise calculating TDM day-by-day without nighttimes (Vallance and Blanc, 2017).

[28] The geographical hierarchies model the interaction among different time series in space, whereas the temporal hierarchies (Yang et al., 2017c) describe the different levels of variability within a single time series. Combining both types of hierarchies may lead to further forecast improvements.

[29] Important note: This performance improvement is computed over all forecast horizons, i.e., 1- to 15-min-ahead. However, at any single horizon, multi-modeling approach performs only marginally better, if not worse, than the best performing component model.

[30] Since PV systems are often installed with a tilt, the inverse transposition algorithms can be used to map the tilted data to a horizontal surface (Killinger et al., 2016; Marion, 2015; Yang et al., 2014c, 2013a).

[31] Clear-sky power output can be expressed as a quadratic model of clear-sky plane-of-

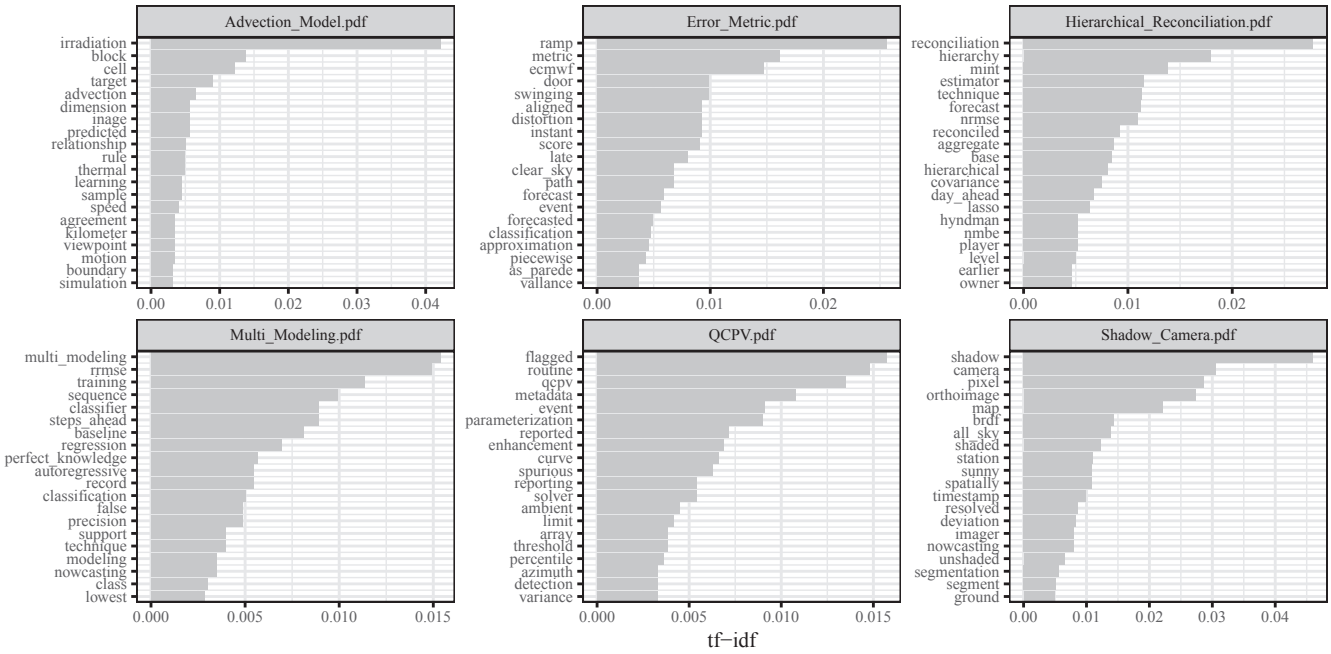

Fig. 6. Top 20 unigrams from each emerging technology, ranked based on term frequency–inverse document frequency (tf–idf).

between the clear-sky power output and the clear-sky plane-of-array irradiance. In the second step, the available PV power output data is controlled using both system-specific analysis (within the system itself) and across-systems analysis. The system-specific analysis can be further divided into smaller steps, such as checking the physical limitation or the daily energy ratio. In parallel, the across-systems analysis considers the effects of cloud enhancement events and comparisons among peers. Since the entire QCPV procedure combines many earlier important contributions to PV data quality control, it can also be regarded as a hybrid method. It is noted that this paper won the best 2016–2017 paper on the topic of solar resources & meteorology, presented by Solar Energy.

6.1.6. Shadow cameras (Kuhn et al., 2017)

Ever since the first major paper[32] on sky-imager-based forecasting was published (Chow et al., 2011), the method has gained recognition in short-term solar forecasting and spawned a large body of literature. However, despite many excellent works in determining the cloud base height (Wang et al., 2016a; Peng et al., 2015; Nguyen and Kleissl, 2014), errors in geolocating 3D cloud objects affects the accuracy of ray tracing, and thus forecast accuracy (Kurtz et al., 2017). Whereas sky imagers take a bottom-up approach in generating irradiance maps over an area, the recently emerged shadow cameras (Kuhn et al., 2017) consider a top-down approach, by taking photos of the ground from an elevated position below the clouds (87 m above ground). Shadow cameras bypass cloud geolocation, thus providing great potential of improvement for short-term solar forecasting.

Shadow cameras generate irradiance maps for two main purposes: (1) provide a benchmark map for sky-imager-based nowcasting systems, which is otherwise unachievable using a few ground sensors; and (2) construct stand-alone nowcasting systems. The shadow camera system developed by Kuhn et al. (2017) uses six cameras to generate an orthonormalized image (orthoimage). By comparing the current orthoimage to an orthoimage taken during a sunny period, the shaded/unshaded/excluded areas can be identified. DNI over the area is subsequently calculated according to the shaded or unshaded pixel. With additional ground-based DHI measurements, GHI maps can be constructed. At present, the forecasting aspects of shadow camera applications are still hypothetical (Kuhn et al., 2017). Nevertheless, its clear advantages over sky imagers (see Section 5 of Kuhn et al., 2017) motivate further studies.

6.2. Analyzing word frequency

Finding the core concept of a document is a central question in text mining. A simple measure of the importance of a word is its *term frequency*, i.e., how often a word occurs in a document. However, stopwords, as mentioned in Section 3.4, tend to appear in all documents. Other context based words such as “solar”, “irradiance” and “forecast” also may appear in most, if not all, documents. Therefore, a more robust measure of the importance of a word is given by term frequency – inverse document frequency (tf–idf). The inverse document frequency (idf) acts as a weight to term frequency; it increases the importance (large weight) of words that appear less often and decreases the importance (small weight) of common words to a set of documents:

$$\mathrm{idf}(x) = \ln\left(\frac{\#\text{ of documents}}{\#\text{ of documents containing the word } x}\right). \tag{2}$$

The measure tf–idf is then given by the product of idf and term frequency.

To analyze the keywords in the six emerging technologies, the top 20 words (unigrams) are selected from each document based on tf–idf. The results are shown in Fig. 6.[33] Based on the earlier short description of each emerging technology, it is obvious that these keywords are relevant to, and expressive about, each document. For example, the list of keywords for “Hierarchical_Reconciliation.pdf” (Yang et al., 2017b) includes “reconciliation”, “hierarchy”, “reconciled” and “hierarchical”,

(*footnote continued*)
array irradiance and metadata (Killinger et al., 2016). Instead of performing the “normal” regression and determining the coefficients, the regression coefficients are taken from previous results; the unknown metadata are treated as the “coefficients” to be fitted.

[32] There are earlier papers (e.g., Nova et al., 2005), but this one is the earliest high-impact paper on this topic.

[33] The words listed in the figure contain abbreviations such as nRMSE, MPIV or ECMWF. Due to plotting space constraints, they are not explicitly expressed in their long forms. However, they are detected during the text processing using the algorithm discussed in Section 5.

which are directly related to the document name. Keywords "mint" (minimum trace), "lasso", "level", "aggregate", "base"[34] and "covariance" describe the technical aspect of the reconciliation method. Other keywords including "day-ahead", "california", "transmission", "state", "nrmse" and "nmbe" reveal the empirical part of the work, namely, a day-ahead forecast reconciliation exercise on transmission level using data from the state of California, evaluated using nRMSE and nMBE. Lastly, keyword "hyndman", as in Rob Hyndman, is the name of an important contributor to the forecast reconciliation technique.

Keywords identified in this manner highlight the content of a document. They thus help the reader to gain quick access to the core contributions of a paper. Although some of the keywords may be difficult to understand based on their literal meaning, knowing them prior to detailed reading is beneficial. Alternatively, searching the keywords from the PDF document itself is also believed to improve efficiency in understanding the concepts.

6.3. Analyzing relationships between words

In addition to knowing the important unigrams, understanding the relationship between words is also meaningful in text mining. This is because: (1) many concepts take more than a word to describe, and (2) a frequently-appearing word may be associated with several other words (e.g., cloud motion, cloud speed, cloud pixel or cloud image). In particular, this section discusses several ways of visualizing the co-occurrence of words. Words can either appear with immediate adjacency, or co-occur within a same paragraph or a sentence. For the first case, a relationship between words can be extracted using n-grams. For the second case, pairwise correlation can be used to examine how often a pair of words appears together relatively to their individual appearances.

Fig. 7 shows the top 20 bigrams extracted from each selected paper on emerging technology. The ranking of bigrams is based on tf–idf. The immediate conclusion that can be drawn from Fig. 7 is that the bigrams reveal additional terms containing words not listed as top unigrams, e.g., neither word from "forecasting skill" in "Advection_Model.pdf" (Inage, 2017) appeared as a top unigram. In other circumstances, words appearing as top unigrams are further illustrated by bigrams, e.g., the word "door" in "Error_Metric.pdf" (Vallance et al., 2017) appears mostly when the swinging-door algorithm—a method to detect ramp events in irradiance time series—is mentioned. This analysis can be extended to n-grams.

Aside from plotting the top bigrams based on descending tf–idf, it is also possible to visualize all bigrams together, as well as their respective counts. Fig. 8 shows a bigram network generated from the documents. It is noted that only bigrams that have more than seven appearances are plotted due to space constraints. The frequency of appearance of a bigram is indicated by the opacity of the arrow linking the words—a darker arrow corresponds to more appearances. In contrast to Fig. 7, the bigram network visualizes the total appearances of each bigram in all documents. This plot is useful in situations where the overall popularity of a concept is of interest. It is observed that several bigrams including "time → series", "power → output" and "camera → system" have high counts, indicating the overall importance of these bigrams in emerging technologies of solar forecasting.[35]

Since the n-grams analysis only examines the relationship between adjacent words, a correlation analysis is used to study the relationship between words located in close proximity. In this paper, a sentence is used to define this proximity. Since the occurrence of a word is binary, i.e., either it is in a sentence or it is not, the *phi coefficient* is used to measure the correlation between two words X and Y:

$$\phi = \frac{n_{11}n_{00} - n_{01}n_{10}}{\sqrt{n_{1\cdot}n_{\cdot 1}n_{0\cdot}n_{\cdot 0}}}, \tag{3}$$

where n_{11} is the number of sentences containing both X and Y; n_{00} is the number of sentences containing neither X nor Y; n_{10} and n_{01} are the numbers of sentences containing one word but not the other; $n_{1\cdot} = n_{11} + n_{10}$; $n_{\cdot 1} = n_{01} + n_{11}$; $n_{0\cdot} = n_{00} + n_{01}$; and $n_{\cdot 0} = n_{10} + n_{00}$. Using the phi coefficient, correlations between all pairs of words in ~2000 sentences from the six documents are computed. There are several ways to analyze the results. One possibility is to filter out word pairs that have a correlation lower than a certain value, and only study highly-correlated pairs. However, similarly to the property of term frequency, highly-correlated pairs do not necessarily indicate the importance of the words. Instead, a high correlation only shows that the words tend to appear together. Alternatively, the phi coefficient results can be used to study the correlating words to a word of interest. As an example, Fig. 9 shows the top 50 words that correlate with "cloud" or "forecast". It is evident that cloud motion, shape, velocity, speed, deformation, etc, often appear in close proximity, indicating various interests when clouds are studied during solar forecasting. Similarly, different forecast methods and applications can be found, e.g., forecast reconciliation, forecast horizon, ECMWF forecasts, forecast error, operational forecast, etc.

6.4. Topic modeling

Topic modeling is a class of statistical approaches that aims at finding unobserved "topics" in text documents. Its simplest application in the present context would be to decide whether a paper returned by a keyword search is really a solar forecasting paper. Topic modeling is actually one of the most important text mining tools. Among various models, latent Dirichlet allocation (LDA) is particularly popular. Its development can be traced back to the works of Blei et al. (2003) and Pritchard et al. (2000), which have a combined citation number of over 40,000. The core idea of LDA is to treat each document as a mixture of topics, where each topic is described by a small amount of words. Therefore, LDA is essential to discover hidden semantic structures in texts. In this preliminary text mining paper, LDA is used for the following purposes: (1) to find out how many topics there are in the six emerging technologies; and (2) to identify overlaps among these technologies.

Due to the unsupervised nature of LDA, the number of topics is either known *a priori*, or chosen based on some metrics—this is very similar to choosing k in the k-means algorithm. It is apparent that the forecasting aspects of the six emerging technologies are distinct, i.e., it is believed that six topics should be set. To verify that, four metrics proposed by Arun et al. (2010), Cao et al. (2009), Deveaud et al. (2014), Griffiths and Steyvers (2004) are used. The reader is referred to the original publications for details on the quantity each metric minimizes or maximizes. By setting the topic numbers from 2 to 15 and running the LDA implementation from the **R** package called **topicmodels**,[36] the corresponding values of the metrics are plotted in Fig. 10. It is evident that having six topics is indeed the optimal choice.

The LDA results using six topics are shown in Fig. 11. Beta on the abscissa is the per-topic-per-word probability. It is clear that words in topics 1 to 6 largely reflect "Shadow_Camera.pdf", "Advection_Model", "Multi_Modeling.pdf", "QCPV.pdf", "Error_Metric.pdf" and "Hierarchical_Forecasting.pdf", respectively. It is noted that Fig. 11 only displays partial LDA results. In fact, LDA assigns a probability to each word in each document. Therefore, to study the commonality in topics, words with small probabilities are filtered out.[37] After examining the

[34] As used in "base forecasts", i.e., forecasts made by various methods before reconciliation.

[35] Words or n-grams having high importance here do not necessarily mean they are involved in these emerging technologies. For example, bigram "time series" might have appeared often as a benchmarking method.

[36] Only 1000 words with highest tf–idf from each document are used.

[37] A word may have high probability in one topic, but not in others. Therefore, after

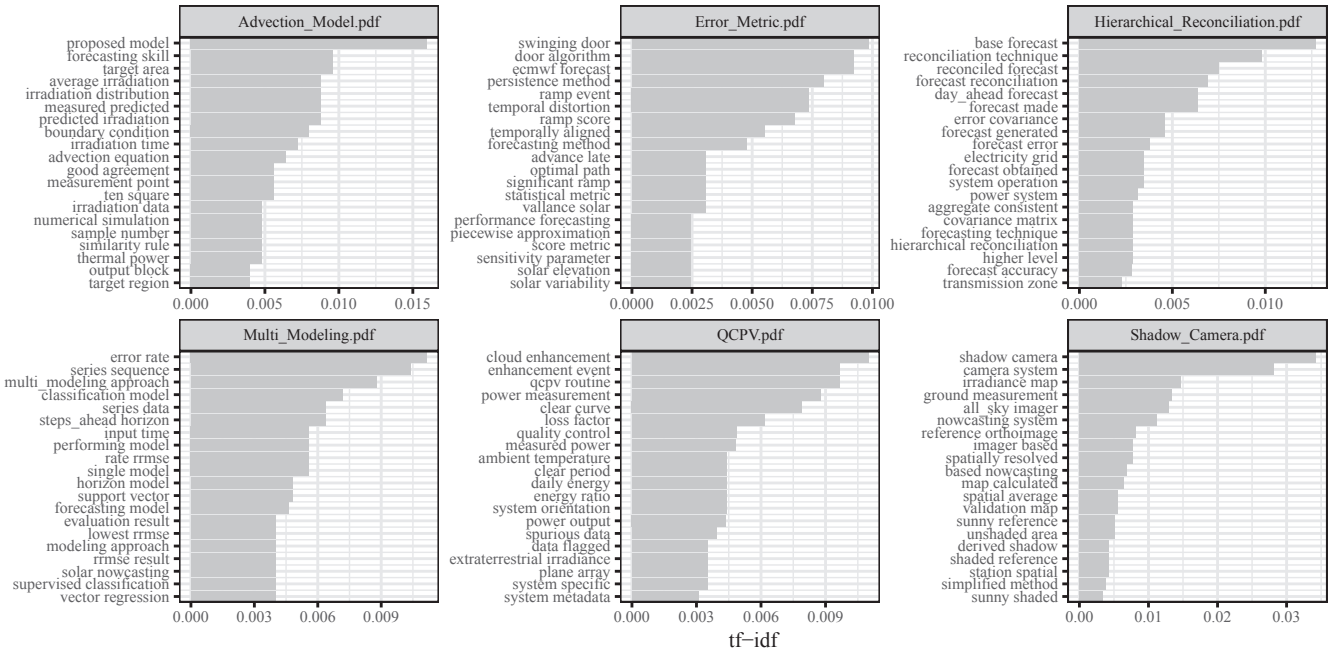

Fig. 7. Top 20 bigrams from each emerging technology, ranked based on tf–idf.

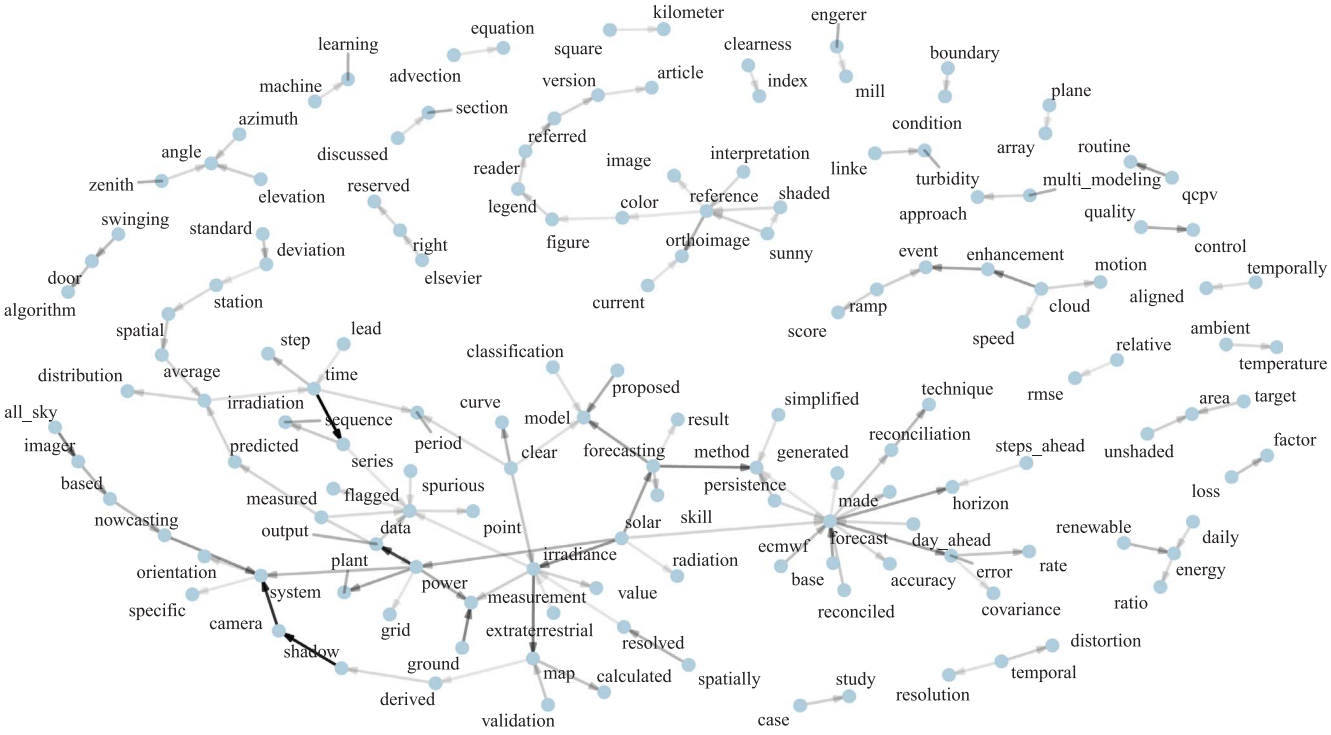

Fig. 8. Bigram network for the emerging technologies. Only bigrams which appear more than seven times are plotted due to space constraints.

complete results—see supplementary material—around 100 words are found to appear in more than one document. Some of these words such as “clear”, “fluctuation”, “station”, “stochastic”, “classification” and “nowcasting” reveal important concerns in solar forecasting. A discussion is presented next.

6.5. Discussion on the emerging technologies based on the text mining results

Although the information shown in Figs. 6–11 is extracted from only the six emerging technologies considered here, it nevertheless suggests important future trends for solar forecasting research.

6.5.1. Future trends on error evaluation and forecast comparison

Using only a single or a few error metrics most likely results in biased opinions on forecasting performance. A suite of metrics should be used instead. Besides the conventional metrics such as nRMSE, nMBE, forecast skill, or KSI, new metrics such as ramp score or TDM

(*footnote continued*)
the filtering process, some words only appear in certain topics. See code output from Appendix A for details.

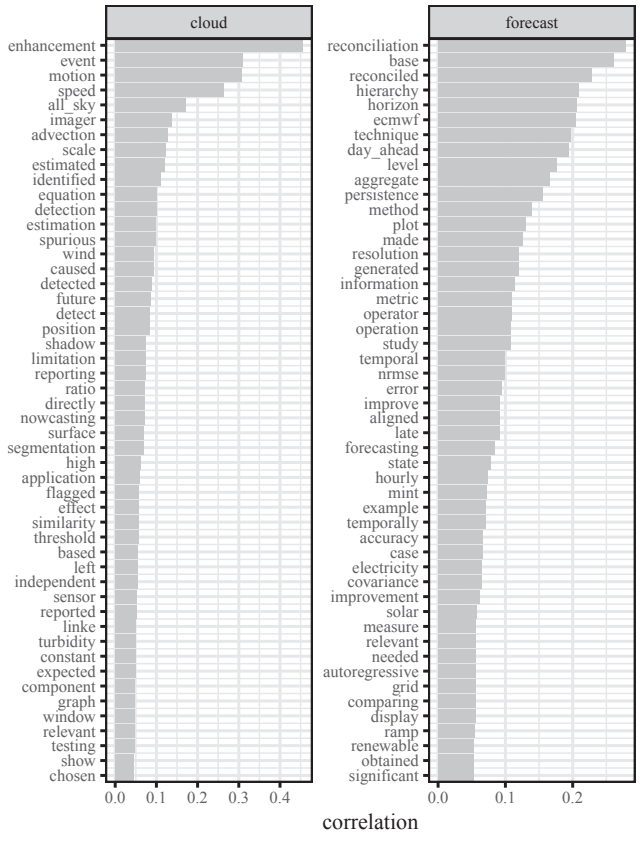

Fig. 9. Top 50 words that correlate with the words "cloud" or "forecast".

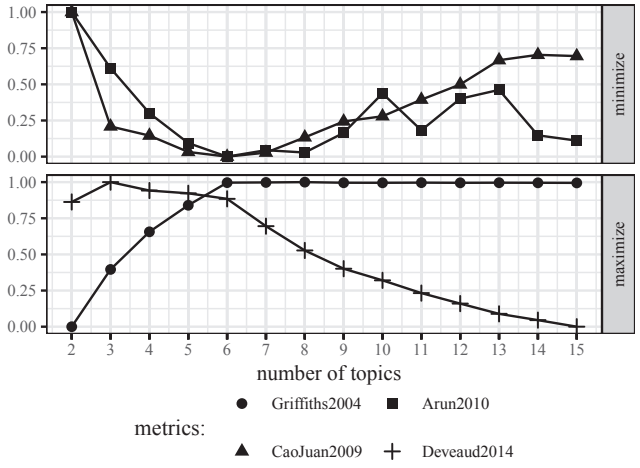

Fig. 10. Four metrics used to estimate the best-fitting number of LDA topics.

(see the important footnote in Section 6.1.2), which capture the unique properties of solar forecasts, should be used whenever possible. Considering that the number of solar forecasting studies is expanding fast, standardizing the error evaluation process would help readers truly compare performance. The lack of common data sources is another main reason hindering forecast comparison. Evaluating forecasts based on standardized datasets (e.g., Hong et al., 2016) is also beneficial to the community in general.

One of the main concerns of having standardized datasets is that many sources of funding are region-specific and applications-oriented; spending time and resources on other datasets that do not have a direct impact may not be of interest. Furthermore, the concern on "unidirectional research"—only forecasts with better accuracies will be accepted by journals—may also lead to potential resistance to standardized datasets. Therefore, the main push towards standardization might originate from research leaders of a smaller community. For example, the artificial intelligence community in solar forecasting might be tempted to poll ideas and develop policies.

6.5.2. Future trends on combining and adjusting forecasts

Forecasts generated by a single method have apparent limitations. Empirical evidence has also shown that combining and adjusting forecasts improve accuracy. In fact, many terms in the literature describe techniques for combining and adjusting forecasts: for example, ensemble forecasting (Sperati et al., 2016), ensemble Kalman filter (Takeda, 2017), ensemble learning (Jiang et al., 2017), multi-modeling (Sanfilippo et al., 2016), reforecast (Chu et al., 2015c), reconciliation (Yang et al., 2017b), MOS (Verzijlbergh et al., 2015), etc. Although the above-mentioned approaches are very different in general, they nevertheless share a same underlying principle: utilizing the strength of each individual model in a smart way, so that an overall better accuracy can be achieved. It is however noted that combining and adjusting forecasts are different from a hybrid model, which uses two or more models in various steps to develop a single forecasting model.

It is observed that ensemble forecasting has been well-established in the NWP solar forecasting community (e.g., Liu et al., 2016; Sperati et al., 2016; Thorey et al., 2015; Zamo et al., 2014a,b). However, despite several attempts, most ensemble learning technologies in the machine learning community have not been transferred to solar forecasting. In this regard, massively applying ensemble learning methods such as bagging, boosting, randomization, option trees, and in particular, *stacking*,[38] is expected to benefit solar forecasting. In addition, statistical ensemble methods, such as additive regression, should also be studied.

6.5.3. Future trends in camera-based forecasting

Improved machine learning techniques both for time series and images will also benefit camera-based forecasting. In addition, tomography applied to concurrent imagery from multiple sites will allow the 3D reconstruction of cloud shapes and cloud optical depth. Kurtz et al. (2017) show that perspective issues contribute to the majority of forecast errors when a single imager is used and the 3-dimensionality of clouds is neglected. Data from multiple high-quality cameras that is being collected in close vicinity at both the Plataforma Solar de Almeria (PSA) and at the University of California (Mejia et al., in preparation) will certainly spawn further research on 3D cloud reconstruction and demonstrate the improvements in forecast accuracy. At PSA, shadow camera images can constrain cloudy voxels in tomography and reduce the number of required upward-looking cameras.

The increasing integration of 3D physics-based models for radiative transfer and fluid motion with sky imagery will contribute to aligning sky observations with real physical phenomena such as thermals, condensation and evaporation, and 3D scattering of solar radiation. Large eddy simulation (LES) captures the 3D wind, moisture, and temperature fields, and has frequently been applied to the study of cloud dynamics. Recently, LES has also been applied to weather forecasting (Schalkwijk et al., 2015). Sky-imager observations can provide some of the initial conditions or assimilation data for LES. Other remote sensors, such as wind lidars, could provide additional information to constrain the flow field, albeit at a considerable expense. 3D radiative transfer is well-established, and could thus be used to compute diffuse radiances throughout the domain for the current composite image from tomography, or forecast cloud fields from LES. To enable such a virtual-reality paradigm, significant computational advances are necessary to compile forecast results within minutes.

[38] Kaggle (https://www.kaggle.com/) is a "playground" for the machine learning community.

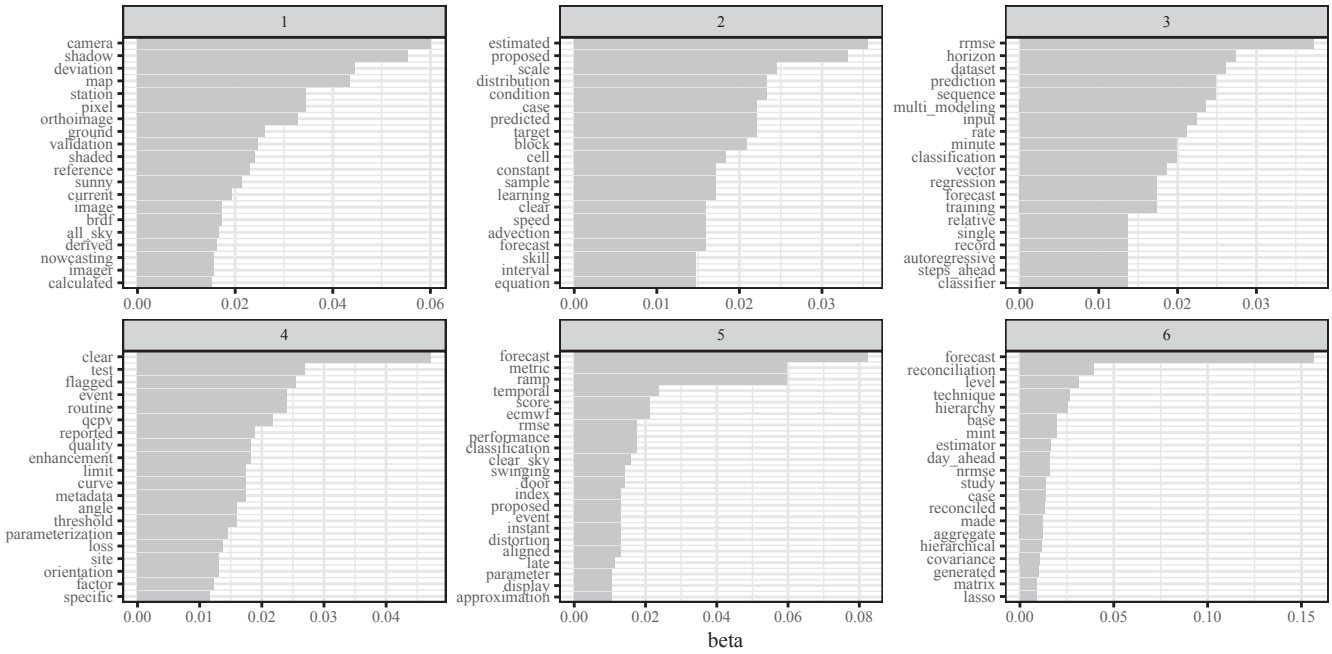

Fig. 11. Top 20 words in each topic selected by Latent Dirichlet Allocation across the six emerging technologies.

6.5.4. Future trends in sensor network-based forecasting

It is now well-known that camera-based forecasts (< 20 min) and NWP (day-ahead) dominantly occupy the two ends of the "line" representing the forecast horizon, whereas for hourly and intra-day forecasts, satellite-based and machine-learning methods are often used. Since clouds are the main source of solar variability, cameras and satellite images have been used as the most appropriate sources of data. However, due the uncertainties embedded in the cloud-to-irradiance conversion, especially with cameras, forecasting solar irradiance from images is still strongly limited by image-to-irradiance translation errors. With more and more distributed PV systems being built, and considering the growing needs for ground truth, sensor networks are believed to play an increasingly important role in future solar forecasting, especially for short-term forecasts.

Central to sensor network-based forecasting are the spatio-temporal dynamical components of the irradiance field. Regressions, spatio-temporal kriging, and PDE approaches are the state-of-the-art methods. Estimation of covariance or, equivalently, correlation function is a key step in spatio-temporal kriging. Currently, all correlation functions used for irradiance kriging are positive functions. However, as demonstrated numerous times by Perez and Fthenakis (2015), Arias-Castro et al. (2014), Lonij et al. (2013), negative correlations are often observed in data at all time-scales. Therefore, it is challenging to correctly model these negative correlations, and at the same time satisfy the statistical properties of correlation functions (e.g., compactly supported, positive definite, stationarity). The hole effect model (Gneiting, 2002) has been known for a long time in the geostatistics community. Studying the hole effect model and other works of Tilmann Gneiting may lead to a breakthrough in solar variability studies, and may eventually lead to forecast improvements.

The parallel method to kriging, namely, PDE, also requires future attention from solar forecasters. It has been shown that using an advection equation on interpolated sensor network data leads to dramatic forecast improvements—a forecast skill of up to 0.8 (Inage, 2017), one of the highest ever reported. Although only a small dataset was used in that study, it is expected that using PDEs on a properly designed sensor network would achieve overall higher forecast accuracies.

6.5.5. Future trends in sensing technologies and data quality

Generating good forecasts requires suitable data—quality-controlled data from multiple sources. Having such data is essential to improve the state-of-the-art forecasting accuracies, which are often limited by the single source of data. Among various sensing technologies mentioned earlier, satellite imagery and ground-based sensor network form a unique pair (e.g., Arbizu-Barrena et al., 2017; Lorenzo et al., 2017). They align perfectly with a well-defined topic in statistics: *integrating low- and high-accuracy experiments* (Zhang et al., 2013; Xiong et al., 2013; Qian and Wu, 2008). Briefly stated, this kind of study is justified whenever many experiments and measurements are available in low accuracy, whereas their high-accuracy counterparts are few. For instance, satellite-derived irradiance datasets provide spatial diversity, but their accuracy is low in general. In parallel, it would be too costly to set up pyranometers everywhere, even though their observations are highly accurate in general. In solar resource assessment, integrating the two data sources is known as site adaptation (Polo et al., 2016). In the future, it would be interesting to consider more rigorous statistical approaches to integrate satellite and ground-based measurements. Similarly, more research is needed in integrating other solar engineering data of this kind. For example, the temporal resolution and accuracy of PV system power output data can be improved by adding observations from several nearby irradiance sensors.

In terms of data quality control, there is no optimal sequence (Gueymard and Ruiz-Arias, 2016). A particular quality issue in one dataset may not exist in another dataset. Therefore, two aspects are critically important here: (1) methods to visualize the data, and (2) various steps available in the literature to handle similar problems. Visualizing data is a scientific subject in itself. For example, there are various ways to visualize a single time series, a moderate number of time series, or a large number of time series (Yang et al., 2017a, 2015b). The reader is referred to the articles published in IEEE Transactions on Visualization and Computer Graphics for visualization techniques. Once a class of quality issue is identified through visualization, the existing control sequence should be considered first. In solar engineering, abundant previous works on quality control of various data types, e.g., sky imager (Schmidt et al., 2016), satellite data (Urraca et al., 2017), PV data (Killinger et al., 2017), irradiance data on a horizontal surface (Gueymard and Ruiz-Arias, 2016) or tilted surfaces (Yang, 2016), are available. It is very likely that any issue in the dataset under scrutiny has been mentioned and handled in previous contributions. Nevertheless, the interactions among various quality control steps should be

noted, since an early filter may affect several later steps.

7. Conclusions

Text mining, as a combination of data science, machine learning, natural language processing, information retrieval and knowledge management, has great potential in transforming how researchers perform literature reviews. Although it has been demonstrated that text mining is able to retrieve and construct technology infrastructures, extract keywords and abbreviations, analyze relationships between words, and perform topic modeling, the study herein presented is still far from what text mining is capable of. For such reasons, the word "preliminary" is added to the title of this paper, and elaborate subsequent contributions are expected.

The foremost drawback of the current research is the amount of data. Despite the fact that 1000 results are analyzed for technology infrastructure, the number is still small in comparison to the total available papers in the literature. In order to make the constructed technology infrastructure more reliable, the search period should be expanded to all years instead of searching within the latest five years. More importantly, to circumvent the limitation of Google Scholar (recall that only the top 1000 results are returned with each search), a nested search can be performed. In other words, once the results from the initial search are analyzed, further searches can be performed based on these results, e.g., the term "imager and camera-based solar forecasting" can be considered, since the initial results provide relatively few papers on this subject. Similar data expansion can be considered for materials presented in other sections of this paper.

With respect to the latter points listed in Section 2.3, it is also of interest to plan such studies in the future. A particularly interesting study would be to design an automatic way of finding out the various ways researchers are contributing to a field, such as the list shown in Section 5.2.4. This is thought feasible if keyword search and topic modeling are combined. By searching the relevant keywords such as "novel" and "new", and performing topic modeling using the surrounding texts, these contents would be summarized into n-grams representing the innovations contained in that paper. A further step to this research is to construct a list of contributions in a chronological order. A potential visualization of such results is the "time river" plot,[39] with which the development of multiple types of contributions can be displayed at once. Lastly, some numerical modeling techniques can be combined with text mining results to form quantified answers to the questions of interest.

Conflict of interest

Authors declare no conflict of interest.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.solener.2017.11.023.

References

Acuna, D.E., Allesina, S., Kording, K.P., 2012. Future impact: Predicting scientific success. Nature 489, 201–202. http://dx.doi.org/10.1038/489201a.

Ahmad, A., Anderson, T.N., Lie, T.T., 2015. Hourly global solar irradiation forecasting for New Zealand. Solar Energy 122, 1398–1408. http://dx.doi.org/10.1016/j.solener.2015.10.055. URL <http://www.sciencedirect.com/science/article/pii/S0038092X15006118> .

Akarslan, E., Hocaoğlu, F.O., 2016. A novel adaptive approach for hourly solar radiation forecasting. Renew. Energy 87 (Part 1), 628–633. http://dx.doi.org/10.1016/j.renene.2015.10.063. URL <http://www.sciencedirect.com/science/article/pii/S096014811530416X> .

Alessandrini, S., Monache, L.D., Sperati, S., Cervone, G., 2015. An analog ensemble for short-term probabilistic solar power forecast. Appl. Energy 157, 95–110. http://dx.doi.org/10.1016/j.apenergy.2015.08.011. URL <http://www.sciencedirect.com/science/article/pii/S0306261915009368> .

Alexandridis, A.K., Zapranis, A.D., 2013. Wavelet neural networks: A practical guide. Neural Netw. 42, 1–27. http://dx.doi.org/10.1016/j.neunet.2013.01.008. URL <http://www.sciencedirect.com/science/article/pii/S0893608013000129> .

Almeida, M.P., Muñoz, M., de la Parra, I., Perpiñán, O., 2017. Comparative study of PV power forecast using parametric and nonparametric PV models. Solar Energy 155, 854–866. http://dx.doi.org/10.1016/j.solener.2017.07.032. URL <http://www.sciencedirect.com/science/article/pii/S0038092X17306175> .

Almeida, M.P., Perpiñán, O., Narvarte, L., 2015. PV power forecast using a nonparametric PV model. Solar Energy 115, 354–368. http://dx.doi.org/10.1016/j.solener.2015.03.006. URL <http://www.sciencedirect.com/science/article/pii/S0038092X15001218> .

Alvo, M., Yu, P.L.H., 2014. Exploratory analysis of ranking data. In: Statistical Methods for Ranking Data. Frontiers in Probability and the Statistical Sciences. Springer, New York, pp. 7–21.

André, M., Dabo-Niang, S., Soubdhan, T., Ould-Baba, H., 2016. Predictive spatio-temporal model for spatially sparse global solar radiation data. Energy 111, 599–608. http://dx.doi.org/10.1016/j.energy.2016.06.004. URL <http://www.sciencedirect.com/science/article/pii/S0360544216307769> .

Antonanzas, J., Osorio, N., Escobar, R., Urraca, R., de Pison, F.M., Antonanzas-Torres, F., 2016. Review of photovoltaic power forecasting. Solar Energy 136, 78–111. http://dx.doi.org/10.1016/j.solener.2016.06.069. URL <http://www.sciencedirect.com/science/article/pii/S0038092X1630250X> .

Arbizu-Barrena, C., Ruiz-Arias, J.A., Rodríguez-Benítez, F.J., Pozo-Vázquez, D., Tovar-Pescador, J., 2017. Short-term solar radiation forecasting by advecting and diffusing MSG cloud index. Solar Energy 155, 1092–1103. http://dx.doi.org/10.1016/j.solener.2017.07.045. URL <http://www.sciencedirect.com/science/article/pii/S0038092X17306308> .

Arias-Castro, E., Kleissl, J., Lave, M., 2014. A Poisson model for anisotropic solar ramp rate correlations. Solar Energy 101, 192–202. http://dx.doi.org/10.1016/j.solener.2013.12.028. URL <http://www.sciencedirect.com/science/article/pii/S0038092X13005549> .

Arun, R., Suresh, V., Veni Madhavan, C.E., Narasimha Murthy, M.N., 2010. On finding the natural number of topics with Latent Dirichlet Allocation: Some observations. In: Zaki, M.J., Yu, J.X., Ravindran, B., Pudi, V. (Eds.), Advances in Knowledge Discovery and Data Mining: 14th Pacific-Asia Conference, PAKDD 2010, Hyderabad, India, June 21–24, 2010. Proceedings. Part I. Springer, Berlin, Heidelberg, pp. 391–402. http://dx.doi.org/10.1007/978-3-642-13657-3_43.

Aryaputera, A.W., Yang, D., Walsh, W.M., 2015a. Day-ahead solar irradiance forecasting in a tropical environment. J. Solar Energy Eng. 137, 051009. http://dx.doi.org/10.1115/1.4030231.

Aryaputera, A.W., Yang, D., Zhao, L., Walsh, W.M., 2015b. Very short-term irradiance forecasting at unobserved locations using spatio-temporal kriging. Solar Energy 122, 1266–1278. http://dx.doi.org/10.1016/j.solener.2015.10.023. URL <http://www.sciencedirect.com/science/article/pii/S0038092X15005745> .

Athanasopoulos, G., Ahmed, R.A., Hyndman, R.J., 2009. Hierarchical forecasts for Australian domestic tourism. Int. J. Forecast. 25, 146–166.

Aybar-Ruiz, A., Jiménez-Fernández, S., Cornejo-Bueno, L., Casanova-Mateo, C., Sanz-Justo, J., Salvador-González, P., Salcedo-Sanz, S., 2016. A novel grouping genetic algorithm – extreme learning machine approach for global solar radiation prediction from numerical weather models inputs. Solar Energy 132, 129–142. http://dx.doi.org/10.1016/j.solener.2016.03.015. URL <http://www.sciencedirect.com/science/article/pii/S0038092X16001985> .

Bacher, P., Madsen, H., Nielsen, H.A., 2009. Online short-term solar power forecasting. Solar Energy 83, 1772–1783. http://dx.doi.org/10.1016/j.solener.2009.05.016. URL <http://www.sciencedirect.com/science/article/pii/S0038092X09001364> .

Barbieri, F., Rajakaruna, S., Ghosh, A., 2017. Very short-term photovoltaic power forecasting with cloud modeling: A review. Renew. Sustain. Energy Rev. 75, 242–263. http://dx.doi.org/10.1016/j.rser.2016.10.068. URL <http://www.sciencedirect.com/science/article/pii/S136403211630733X> .

Bartholomy, O., Vargas, T., Simone, M., Hansen, C., Fitchett, S., Pohl, A., 2014. Benchmarking solar power and irradiance forecasting accuracy at Sacramento Municipal Utility District. In: 2014 IEEE 40th Photovoltaic Specialist Conference (PVSC), pp. 63–68, doi:http://dx.doi.org/10.1109/PVSC.2014.6925196.

Bernecker, D., Riess, C., Angelopoulou, E., Hornegger, J., 2014. Continuous short-term irradiance forecasts using sky images. Solar Energy 110, 303–315. http://dx.doi.org/10.1016/j.solener.2014.09.005. URL <http://www.sciencedirect.com/science/article/pii/S0038092X14004356> .

Bessa, R., Trindade, A., Silva, C.S., Miranda, V., 2015. Probabilistic solar power forecasting in smart grids using distributed information. Int. J. Electr. Power Energy Syst. 72, 16–23. http://dx.doi.org/10.1016/j.ijepes.2015.02.006. URL, the Special Issue for 18th Power Systems Computation Conference <http://www.sciencedirect.com/science/article/pii/S0142061515000897> .

Beyer, H.G., Polo Martinez, J., Suri, M., Torres, J.L., Lorenz, E., Müller, S.C., Hoyer-Klick, C., Ineichen, P., 2009. D 1.1. 3 Report on Benchmarking of Radiation Products. Technical Report 038665. Management and Exploitation of Solar Resource Knowledge < http://www.mesor.org/docs/MESoR_Benchmarking_of_radiation_products.pdf > .

Bhatti, M.A., 2000. Practical Methods of Optimization. Springer.

Bigdeli, N., Borujeni, M.S., Afshar, K., 2017. Time series analysis and short-term forecasting of solar irradiation, a new hybrid approach. Swarm Evolution. Comput. 34, 75–88. http://dx.doi.org/10.1016/j.swevo.2016.12.004. URL <http://www.sciencedirect.com/science/article/pii/S2210650216305673> .

[39] https://rud.is/b/2016/06/28/making-time-rivers-in-r/.

Blanc, P., Espinar, B., Geuder, N., Gueymard, C., Meyer, R., Pitz-Paal, R., Reinhardt, B., Renné, D., Sengupta, M., Wald, L., Wilbert, S., 2014. Direct normal irradiance related definitions and applications: The circumsolar issue. Solar Energy 110, 561–577. http://dx.doi.org/10.1016/j.solener.2014.10.001. URL <http://www.sciencedirect.com/science/article/pii/S0038092X14004824> .

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet Allocation. J. Mach. Learn. Res. 3, 993–1022.

Boata, R.S., Gravila, P., 2012. Functional fuzzy approach for forecasting daily global solar irradiation. Atmos. Res. 112, 79–88. http://dx.doi.org/10.1016/j.atmosres.2012.04.011. URL <http://www.sciencedirect.com/science/article/pii/S0169809512001172> .

Boilley, A., Thomas, C., Marchand, M., Wey, E., Blanc, P., 2016. The solar forecast similarity method: A new method to compute solar radiation forecasts for the next day. Energy Proc. 91, 1018–1023. http://dx.doi.org/10.1016/j.egypro.2016.06.270. URL, proceedings of the 4th International Conference on Solar Heating and Cooling for Buildings and Industry (SHC 2015) <http://www.sciencedirect.com/science/article/pii/S1876610216303708> .

Boland, J., 2015. Spatial-temporal forecasting of solar radiation. Renew. Energy 75, 607–616. http://dx.doi.org/10.1016/j.renene.2014.10.035. URL <http://www.sciencedirect.com/science/article/pii/S0960148114006624> .

Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. J. Economet. 31, 307–327. http://dx.doi.org/10.1016/0304-4076(86)90063-1. URL <http://www.sciencedirect.com/science/article/pii/0304407686900631> .

Bouzerdoum, M., Mellit, A., Pavan, A.M., 2013. A hybrid model (SARIMA–SVM) for short-term power forecasting of a small-scale grid-connected photovoltaic plant. Solar Energy 98 (Part C), 226–235. http://dx.doi.org/10.1016/j.solener.2013.10.002. URL <http://www.sciencedirect.com/science/article/pii/S0038092X13004039> .

Bouzgou, H., Gueymard, C.A., 2017. Minimum redundancy – maximum relevance with extreme learning machines for global solar radiation forecasting: Toward an optimized dimensionality reduction for solar time series. Solar Energy 158, 595–609. http://dx.doi.org/10.1016/j.solener.2017.10.035. URL <http://www.sciencedirect.com/science/article/pii/S0038092X17309052> .

Box, G.E.P., Jenkins, G.M., 1994. Time Series Analysis: Forecasting and Control. third ed. Prentice Hall PTR, Upper Saddle River, NJ, USA.

Brabec, M., Paulescu, M., Badescu, V., 2015. Tailored vs black-box models for forecasting hourly average solar irradiance. Solar Energy 111, 320–331. http://dx.doi.org/10.1016/j.solener.2014.11.003. URL <http://www.sciencedirect.com/science/article/pii/S0038092X14005350> .

Breitkreuz, H., Schroedter-Homscheidt, M., Holzer-Popp, T., 2007. A case study to prepare for the utilization of aerosol forecasts in solar energy industries. Solar Energy 81, 1377–1385. http://dx.doi.org/10.1016/j.solener.2007.01.009. URL <http://www.sciencedirect.com/science/article/pii/S0038092X07000278> .

Breitkreuz, H., Schroedter-Homscheidt, M., Holzer-Popp, T., Dech, S., 2009. Short-range direct and diffuse irradiance forecasts for solar energy applications based on aerosol chemical transport and numerical weather modeling. J. Appl. Meteorol. Climatol. 48, 1766–1779. http://dx.doi.org/10.1175/2009JAMC2090.1.

Brier, G.W., 1950. Verification of forecasts expressed in terms of probability. Mont. Weath. Rev. 78, 1–3.

Cao, J., Lin, X., 2008. Study of hourly and daily solar irradiation forecast using diagonal recurrent wavelet neural networks. Energy Convers. Manage. 49, 1396–1406. http://dx.doi.org/10.1016/j.enconman.2007.12.030. URL <http://www.sciencedirect.com/science/article/pii/S0196890408000125> .

Cao, J., Xia, T., Li, J., Zhang, Y., Tang, S., 2009. A density-based method for adaptive LDA model selection. Neurocomputing 72, 1775–1781.

Cervone, G., Clemente-Harding, L., Alessandrini, S., Monache, L.D., 2017. Short-term photovoltaic power forecasting using artificial neural networks and an analog ensemble. Renew. Energy 108, 274–286. http://dx.doi.org/10.1016/j.renene.2017.02.052. URL <http://www.sciencedirect.com/science/article/pii/S0960148117301386> .

Chen, C., Duan, S., Cai, T., Liu, B., 2011. Online 24-h solar power forecasting based on weather type classification using artificial neural network. Solar Energy 85, 2856–2870. http://dx.doi.org/10.1016/j.solener.2011.08.027. URL <http://www.sciencedirect.com/science/article/pii/S0038092X11003008> .

Chen, S.X., Gooi, H.B., Wang, M.Q., 2013. Solar radiation forecast based on fuzzy logic and neural networks. Renew. Energy 60, 195–201. http://dx.doi.org/10.1016/j.renene.2013.05.011. URL <http://www.sciencedirect.com/science/article/pii/S0960148113002565> .

Chen, Y., Ding, C., Hu, J., Chen, R., Hui, P., Fu, X., 2017. Building and analyzing a global co-authorship network using Google Scholar data. In: Proceedings of the 26th International Conference on World Wide Web Companion, International World Wide Web Conferences Steering Committee, pp. 1219–1224, doi:http://dx.doi.org/10.1145/3041021.3053056.

Chow, C.W., Belongie, S., Kleissl, J., 2015. Cloud motion and stability estimation for intra-hour solar forecasting. Solar Energy 115, 645–655. http://dx.doi.org/10.1016/j.solener.2015.03.030. URL <http://www.sciencedirect.com/science/article/pii/S0038092X15001565> .

Chow, C.W., Urquhart, B., Lave, M., Dominguez, A., Kleissl, J., Shields, J., Washom, B., 2011. Intra-hour forecasting with a total sky imager at the UC San Diego solar energy testbed. Solar Energy 85, 2881–2893. http://dx.doi.org/10.1016/j.solener.2011.08.025. URL <http://www.sciencedirect.com/science/article/pii/S0038092X11002982> .

Chu, Y., Coimbra, C.F.M., 2017. Short-term probabilistic forecasts for direct normal irradiance. Renew. Energy 101, 526–536. http://dx.doi.org/10.1016/j.renene.2016.09.012. URL <http://www.sciencedirect.com/science/article/pii/S0960148116308011> .

Chu, Y., Li, M., Coimbra, C.F.M., 2016. Sun-tracking imaging system for intra-hour DNI forecasts. Renew. Energy 96 (Part A), 792–799. http://dx.doi.org/10.1016/j.renene.2016.05.041. URL <http://www.sciencedirect.com/science/article/pii/S0960148116304529> .

Chu, Y., Li, M., Pedro, H.T.C., Coimbra, C.F.M., 2015a. Real-time prediction intervals for intra-hour DNI forecasts. Renew. Energy 83, 234–244. http://dx.doi.org/10.1016/j.renene.2015.04.022. URL <http://www.sciencedirect.com/science/article/pii/S096014811500302X> .

Chu, Y., Pedro, H.T.C., Coimbra, C.F.M., 2013. Hybrid intra-hour DNI forecasts with sky image processing enhanced by stochastic learning. Solar Energy 98 (Part C), 592–603. http://dx.doi.org/10.1016/j.solener.2013.10.020. URL <http://www.sciencedirect.com/science/article/pii/S0038092X13004325> .

Chu, Y., Pedro, H.T.C., Li, M., Coimbra, C.F.M., 2015b. Real-time forecasting of solar irradiance ramps with smart image processing. Solar Energy 114, 91–104. http://dx.doi.org/10.1016/j.solener.2015.01.024. URL <http://www.sciencedirect.com/science/article/pii/S0038092X15000389> .

Chu, Y., Urquhart, B., Gohari, S.M.I., Pedro, H.T.C., Kleissl, J., Coimbra, C.F.M., 2015c. Short-term reforecasting of power output from a 48 MWe solar PV plant. Solar Energy 112, 68–77. http://dx.doi.org/10.1016/j.solener.2014.11.017. URL <http://www.sciencedirect.com/science/article/pii/S0038092X14005611> .

Cox, T.F., Cox, M.A.A., 2000. Multidimensional Scaling. CRC press.

Cressie, N., Wikle, C.K., 2011. Statistics for Spatio-Temporal Data. Wiley Series in Probability and Statistics. Wiley URL <http://books.google.com.sg/books?id=-kOC6D0DiNYC> .

Dagestad, K.F., Olseth, J.A., 2007. A modified algorithm for calculating the cloud index. Solar Energy 81, 280–289. http://dx.doi.org/10.1016/j.solener.2005.12.010. URL <http://www.sciencedirect.com/science/article/pii/S0038092X06000624> .

David, M., Ramahatana, F., Trombe, P., Lauret, P., 2016. Probabilistic forecasting of the solar irradiance with recursive ARMA and GARCH models. Solar Energy 133, 55–72. http://dx.doi.org/10.1016/j.solener.2016.03.064. URL <http://www.sciencedirect.com/science/article/pii/S0038092X16300172> .

Deardorff, J.W., 1972. Parameterization of the planetary boundary layer for use in general circulation models. Mont. Weath. Rev. 100, 93–106. http://dx.doi.org/10.1175/1520-0493(1972)100<0093:POTPBL>2.3.CO;2.

Debnath, L., Shah, F.A., 2015. Wavelet Transforms and Their Applications. Springer.

Dedinec, A., Filiposka, S., Dedinec, A., Kocarev, L., 2016. Deep belief network based electricity load forecasting: An analysis of Macedonian case. Energy 115 (Part 3), 1688–1700. http://dx.doi.org/10.1016/j.energy.2016.07.090. URL, sustainable Development of Energy, Water and Environment Systems <http://www.sciencedirect.com/science/article/pii/S0360544216310076> .

Delen, D., Crossland, M.D., 2008. Seeding the survey and analysis of research literature with text mining. Expert Syst. Appl. 34, 1707–1720. http://dx.doi.org/10.1016/j.eswa.2007.01.035. URL <http://www.sciencedirect.com/science/article/pii/S0957417407000486> .

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Statist. Soc. Ser. B (Methodol.) 39, 1–38. http://dx.doi.org/10.2307/2984875. URL <http://www.jstor.org/stable/2984875> .

Deo, R.C., Şahin, M., 2017. Forecasting long-term global solar radiation with an ANN algorithm coupled with satellite-derived (MODIS) land surface temperature (LST) for regional locations in Queensland. Renew. Sustain. Energy Rev. 72, 828–848. http://dx.doi.org/10.1016/j.rser.2017.01.114. URL <http://www.sciencedirect.com/science/article/pii/S1364032117301247> .

Deo, R.C., Wen, X., Qi, F., 2016. A wavelet-coupled support vector machine model for forecasting global incident solar radiation using limited meteorological dataset. Appl. Energy 168, 568–593. http://dx.doi.org/10.1016/j.apenergy.2016.01.130. URL <http://www.sciencedirect.com/science/article/pii/S0306261916301180> .

Desai, P.S., 2013. Editorial–Marketing Science replication and disclosure policy. Market. Sci. 32, 1–3. http://dx.doi.org/10.1287/mksc.1120.0761.

Deveaud, R., SanJuan, E., Bellot, P., 2014. Accurate and effective latent concept modeling for ad hoc information retrieval. Document numérique 17, 61–84.

Diagne, M., David, M., Lauret, P., Boland, J., Schmutz, N., 2013. Review of solar irradiance forecasting methods and a proposition for small-scale insular grids. Renew. Sustain. Energy Rev. 27, 65–76. http://dx.doi.org/10.1016/j.rser.2013.06.042. URL <http://www.sciencedirect.com/science/article/pii/S1364032113004334> .

Dickey, D.G., 2011. Dickey-Fuller tests. In: Lovric, M. (Ed.), International Encyclopedia of Statistical Science. Springer, Berlin Heidelberg, pp. 385–388. URL <http://link.springer.com/referenceworkentry/10.1007/978-3-642-04898-2_210> .

Diebold, F.X., 2015. Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold–Mariano tests. J. Bus. Econ. Statist. 33, 1–10. http://dx.doi.org/10.1080/07350015.2014.983236.

Diebold, F.X., Mariano, R.S., 1995. Comparing predictive accuracy. J. Bus. Econ. Statist. 13, 134–144. http://dx.doi.org/10.1198/073500102753410444.

Dong, Z., Yang, D., Reindl, T., Walsh, W.M., 2013. Short-term solar irradiance forecasting using exponential smoothing state space model. Energy 55, 1104–1113. http://dx.doi.org/10.1016/j.energy.2013.04.027. URL <http://www.sciencedirect.com/science/article/pii/S0360544213003381> .

Dong, Z., Yang, D., Reindl, T., Walsh, W.M., 2014. Satellite image analysis and a hybrid ESSS/ANN model to forecast solar irradiance in the tropics. Energy Convers. Manage. 79, 66–73. http://dx.doi.org/10.1016/j.enconman.2013.11.043. URL <http://www.sciencedirect.com/science/article/pii/S0196890413007644> .

Dong, Z., Yang, D., Reindl, T., Walsh, W.M., 2015. A novel hybrid approach based on self-organizing maps, support vector regression and particle swarm optimization to forecast solar irradiance. Energy 82, 570–577. http://dx.doi.org/10.1016/j.energy.2015.01.066. URL <http://www.sciencedirect.com/science/article/pii/S0360544215000900> .

ECMWF, 2017. IFS Documentation – CY43R1 < https://www.ecmwf.int/search/elibrary/part?title=part&secondary_title=43R1 > .

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. Ann. Statist. 32, 407–499. http://dx.doi.org/10.1214/009053604000000067. URL <http://projecteuclid.org/euclid.aos/1083178935> .

Engerer, N.A., Mills, F.P., 2014. Kpv: A clear-sky index for photovoltaics. Solar Energy 105, 679–693. http://dx.doi.org/10.1016/j.solener.2014.04.019. URL <http://www.sciencedirect.com/science/article/pii/S0038092X14002151> .

Engle, R.F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. Econometrica 50, 987–1007. URL <http://www.jstor.org/stable/1912773> .

Espinar, B., Ramírez, L., Drews, A., Beyer, H.G., Zarzalejo, L.F., Polo, J., Martín, L., 2009. Analysis of different comparison parameters applied to solar radiation data from satellite and German radiometric stations. Solar Energy 83, 118–125. http://dx.doi.org/10.1016/j.solener.2008.07.009. URL <http://www.sciencedirect.com/science/article/pii/S0038092X08001655> .

Feinerer, I., Hornik, K., 2017. tm: Text Mining Package. < https://CRAN.R-project.org/package=tm > . r package version 0.7-1.

Feinerer, I., Hornik, K., Meyer, D., 2008. Text mining infrastructure in R. J. Statist. Softw. 25, 1–54. URL <http://www.jstatsoft.org/v25/i05/> .

Feldman, R., Sanger, J., 2007. The Text Mining Handbook. Cambridge University Press.

Fernandez-Jimenez, L.A., Muñoz-Jimenez, A., Falces, A., Mendoza-Villena, M., Garcia-Garrido, E., Lara-Santillan, P.M., Zorzano-Alba, E., Zorzano-Santamaria, P.J., 2012. Short-term power forecasting system for photovoltaic plants. Renew. Energy 44, 311–317. http://dx.doi.org/10.1016/j.renene.2012.01.108. URL <http://www.sciencedirect.com/science/article/pii/S0960148112001516> .

Fisher, R.A., 1937. The design of experiments. Oliver And Boyd, Edinburgh; London.

Fletcher, R., 2013. Practical Optimization Methods: With Mathematica Applications. John Wiley & Sons.

Frías-Paredes, L., Mallor, F., León, T., Gastón-Romeo, M., 2016. Introducing the temporal distortion index to perform a bidimensional analysis of renewable energy forecast. Energy 94, 180–194. http://dx.doi.org/10.1016/j.energy.2015.10.093. URL <http://www.sciencedirect.com/science/article/pii/S0360544215014619> .

Gagne II, D.J., McGovern, A., Haupt, S.E., Williams, J.K., 2017. Evaluation of statistical learning configurations for gridded solar irradiance forecasting. Solar Energy 150, 383–393. http://dx.doi.org/10.1016/j.solener.2017.04.031. URL <http://www.sciencedirect.com/science/article/pii/S0038092X17303158> .

Gala, Y., Fernández, Á., Díaz, J., Dorronsoro, J., 2016. Hybrid machine learning forecasting of solar radiation values. Neurocomputing 176, 48–59. http://dx.doi.org/10.1016/j.neucom.2015.02.078. URL <http://www.sciencedirect.com/science/article/pii/S0925231215005536> .

Ghayekhloo, M., Ghofrani, M., Menhaj, M.B., Azimi, R., 2015. A novel clustering approach for short-term solar radiation forecasting. Solar Energy 122, 1371–1383. http://dx.doi.org/10.1016/j.solener.2015.10.053. URL <http://www.sciencedirect.com/science/article/pii/S0038092X1500609X> .

Ghonima, M.S., Urquhart, B., Chow, C.W., Shields, J.E., Cazorla, A., Kleissl, J., 2012. A method for cloud detection and opacity classification based on ground based sky imagery. Atmos. Measur. Techniq. 5, 2881–2892. http://dx.doi.org/10.5194/amt-5-2881-2012. URL <https://www.atmos-meas-tech.net/5/2881/2012/> .

Gneiting, T., 2002. Compactly supported correlation functions. J. Mult. Anal. 83, 493–508. http://dx.doi.org/10.1006/jmva.2001.2056. URL <http://www.sciencedirect.com/science/article/pii/S0047259X01920561> .

Gneiting, T., Raftery, A.E., Westveld III, A.H., Goldman, T., 2005. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Mont. Weath. Rev. 133, 1098–1118. http://dx.doi.org/10.1175/MWR2904.1.

Golub, G.H., Heath, M., Wahba, G., 1979. Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics 21, 215–223. http://dx.doi.org/10.1080/00401706.1979.10489751.

Gooijer, J.G.D., Hyndman, R.J., 2006. 25 years of time series forecasting. Int. J. Forecast. 22, 443–473. http://dx.doi.org/10.1016/j.ijforecast.2006.01.001. URL, twenty five years of forecasting <http://www.sciencedirect.com/science/article/pii/S0169207006000021> .

Gouriéroux, C., Holly, A., Monfort, A., 1982. Likelihood ratio test, Wald test, and Kuhn-Tucker test in linear models with inequality constraints on the regression parameters. Econometrica 50, 63–80. http://dx.doi.org/10.2307/1912529. URL <http://www.jstor.org/stable/1912529> .

Grantham, A., Gel, Y.R., Boland, J., 2016. Nonparametric short-term probabilistic forecasting for solar radiation. Solar Energy 133, 465–475. http://dx.doi.org/10.1016/j.solener.2016.04.011. URL <http://www.sciencedirect.com/science/article/pii/S0038092X16300342> .

Gray, S.F., 1996. Modeling the conditional distribution of interest rates as a regime-switching process. J. Fin. Econ. 42, 27–62.

Griffiths, T.L., Steyvers, M., 2004. Finding scientific topics. Proc. Nat. Acad. Sci. 101, 5228–5235.

Gueymard, C.A., 2008. REST2: High-performance solar radiation model for cloudless-sky irradiance, illuminance, and photosynthetically active radiation - validation with a benchmark dataset. Solar Energy 82, 272–285. http://dx.doi.org/10.1016/j.solener.2007.04.008. URL <http://www.sciencedirect.com/science/article/pii/S0038092X07000990> .

Gueymard, C.A., 2012a. Clear-sky irradiance predictions for solar resource mapping and large-scale applications: Improved validation methodology and detailed performance analysis of 18 broadband radiative models. Solar Energy 86, 2145–2169. http://dx.doi.org/10.1016/j.solener.2011.11.011. URL, progress in Solar Energy 3 <http://www.sciencedirect.com/science/article/pii/S0038092X11004221> .

Gueymard, C.A., 2012b. Temporal variability in direct and global irradiance at various time scales as affected by aerosols. Solar Energy 86, 3544–3553. http://dx.doi.org/10.1016/j.solener.2012.01.013. URL, solar Resources <http://www.sciencedirect.com/science/article/pii/S0038092X12000291> .

Gueymard, C.A., 2014. A review of validation methodologies and statistical performance indicators for modeled solar radiation data: Towards a better bankability of solar projects. Renew. Sustain. Energy Rev. 39, 1024–1034. http://dx.doi.org/10.1016/j.rser.2014.07.117. URL <http://www.sciencedirect.com/science/article/pii/S1364032114005693> .

Gueymard, C.A., 2017. Cloud and albedo enhancement impacts on solar irradiance using high-frequency measurements from thermopile and photodiode radiometers. Part 2: Performance of separation and transposition models for global tilted irradiance. Solar Energy 153, 766–779. http://dx.doi.org/10.1016/j.solener.2017.04.068. URL <http://www.sciencedirect.com/science/article/pii/S0038092X17303730> .

Gueymard, C.A., Myers, D.R., 2008. Solar radiation measurement: Progress in radiometry for improved modeling. In: Badescu, V. (Ed.), Modeling Solar Radiation at the Earth's Surface: Recent Advances. Springer, Berlin, Heidelberg, pp. 1–27. http://dx.doi.org/10.1007/978-3-540-77455-6_1.

Gueymard, C.A., Renné, D., Vignola, F.E., 2009. Editorial: Journal's performance and publication criteria. Solar Energy 83, 1. http://dx.doi.org/10.1016/j.solener.2008.07.007. URL <http://www.sciencedirect.com/science/article/pii/S0038092X08001801> .

Gueymard, C.A., Ruiz-Arias, J.A., 2015. Validation of direct normal irradiance predictions under arid conditions: A review of radiative models and their turbidity-dependent performance. Renew. Sustain. Energy Rev. 45, 379–396. http://dx.doi.org/10.1016/j.rser.2015.01.065. URL <http://www.sciencedirect.com/science/article/pii/S1364032115000751> .

Gueymard, C.A., Ruiz-Arias, J.A., 2016. Extensive worldwide validation and climate sensitivity analysis of direct irradiance predictions from 1-min global irradiance. Solar Energy 128, 1–30. http://dx.doi.org/10.1016/j.solener.2015.10.010. URL, special issue: Progress in Solar Energy <http://www.sciencedirect.com/science/article/pii/S0038092X15005435> .

Gulin, M., Pavlović, T., Vašak, M., 2017. A one-day-ahead photovoltaic array power production prediction with combined static and dynamic on-line correction. Solar Energy 142, 49–60. http://dx.doi.org/10.1016/j.solener.2016.12.008. URL <http://www.sciencedirect.com/science/article/pii/S0038092X16306053> .

Gutierrez-Corea, F.V., Manso-Callejo, M.A., Moreno-Regidor, M.P., Manrique-Sancho, M.T., 2016. Forecasting short-term solar irradiance based on artificial neural networks and data from neighboring meteorological stations. Solar Energy 134, 119–131. http://dx.doi.org/10.1016/j.solener.2016.04.020. URL <http://www.sciencedirect.com/science/article/pii/S0038092X16300536> .

Hamill, T.M., Nehrkorn, T., 1993. A short-term cloud forecast scheme using cross correlations. Weat. Forecast. 8, 401–411. http://dx.doi.org/10.1175/1520-0434(1993)008<0401:ASTCFS>2.0.CO;2.

Hammer, A., Heinemann, D., Hoyer, C., Kuhlemann, R., Lorenz, E., Müller, R., Beyer, H.G., 2003. Solar energy assessment using remote sensing technologies. Rem. Sens. Environ. 86, 423–432. http://dx.doi.org/10.1016/S0034-4257(03)00083-X. URL, urban Remote Sensing <http://www.sciencedirect.com/science/article/pii/S003442570300083X> .

Harzing, A.W., 2013. A preliminary test of Google Scholar as a source for citation data: A longitudinal study of Nobel prize winners. Scientometrics 94, 1057–1075. http://dx.doi.org/10.1007/s11192-012-0777-7.

Hastie, T., Tibshirani, R., 1990. Generalized Additive Models. Wiley Online Library.

Hastie, T., Tibshirani, R., Friedman, J., 2009. Linear methods for regression. In: The Elements of Statistical Learning. Springer Series in Statistics. Springer, New York, pp. 43–99. URL <http://link.springer.com/chapter/10.1007/978-0-387-84858-7_3> .

Haupt, S.E., Kosović, B., 2017. Variable generation power forecasting as a big data problem. IEEE Trans. Sustain. Energy 8, 725–732. http://dx.doi.org/10.1109/TSTE.2016.2604679.

Hersbach, H., 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. Weat. Forecast. 15, 559–570.

Hinton, G.E., Osindero, S., Teh, Y.W., 2006. A fast learning algorithm for deep belief nets. Neural Comput. 18, 1527–1554.

Hirsch, J.E., 2005. An index to quantify an individual's scientific research output. Proc. Nat. Acad. f Sci. U.S.A. 102, 16569–16572. http://dx.doi.org/10.1073/pnas.0507655102. URL <http://www.pnas.org/content/102/46/16569.abstract> .

Hoff, T.E., Perez, R., Kleissl, J., Renne, D., Stein, J., 2013. Reporting of irradiance modeling relative prediction errors. Prog. Photovolt.: Res. Appl. 21, 1514–1519. http://dx.doi.org/10.1002/pip.2225.

Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., Hyndman, R.J., 2016. Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond. Int. J. Forecast. 32, 896–913. http://dx.doi.org/10.1016/j.ijforecast.2016.02.001. URL <http://www.sciencedirect.com/science/article/pii/S0169207016000133> .

Huang, J., Davy, R.J., 2016. Predicting intra-hour variability of solar irradiance using hourly local weather forecasts. Solar Energy 139, 633–639. http://dx.doi.org/10.1016/j.solener.2016.10.036. URL <http://www.sciencedirect.com/science/article/pii/S0038092X16305011> .

Huang, J., Korolkiewicz, M., Agrawal, M., Boland, J., 2013. Forecasting solar radiation on an hourly time scale using a coupled autoregressive and dynamical system (CARDS) model. Solar Energy 87, 136–149. http://dx.doi.org/10.1016/j.solener.2012.10.012. URL <http://www.sciencedirect.com/science/article/pii/S0038092X12003775> .

Huang, J., Thatcher, M., 2017. Assessing the value of simulated regional weather variability in solar forecasting using numerical weather prediction. Solar Energy 144, 529–539. http://dx.doi.org/10.1016/j.solener.2017.01.058. URL <http://www.sciencedirect.com/science/article/pii/S0038092X17300774> .

Hyndman, R.J., Ahmed, R.A., Athanasopoulos, G., Shang, H.L., 2011. Optimal combination forecasts for hierarchical time series. Comput. Statist. Data Anal. 55, 2579–2589.

Hyndman, R.J., Koehler, A.B., Ord, J.K., Snyder, R.D., 2008. Forecasting with Exponential

Smoothing. Springer, Deblik, Berlin, Germany.
Inage, S., 2017. Development of an advection model for solar forecasting based on ground data first report: Development and verification of a fundamental model. Solar Energy 153, 414–434. http://dx.doi.org/10.1016/j.solener.2017.05.019. URL <http://www.sciencedirect.com/science/article/pii/S0038092X17303961> .
Ineichen, P., 2008. A broadband simplified version of the Solis clear sky model. Solar Energy 82, 758–762. http://dx.doi.org/10.1016/j.solener.2008.02.009. URL <http://www.sciencedirect.com/science/article/pii/S0038092X08000406> .
Inman, R.H., Pedro, H.T.C., Coimbra, C.F.M., 2013. Solar forecasting methods for renewable energy integration. Prog. Energy Combust. Sci. 39, 535–576. http://dx.doi.org/10.1016/j.pecs.2013.06.002. URL <http://www.sciencedirect.com/science/article/pii/S0360128513000294> .
Ishwarappa, Anuradha, J., 2015. A brief introduction on big data 5Vs characteristics and Hadoop technology. Proc. Comput. Sci. 48, 319–324. http://dx.doi.org/10.1016/j.procs.2015.04.188. URL, international Conference on Computer, Communication and Convergence (ICCC 2015) <http://www.sciencedirect.com/science/article/pii/S1877050915006973> .
Jamaly, M., Kleissl, J., 2017. Spatiotemporal interpolation and forecast of irradiance data using kriging. Solar Energy 158, 407–423. http://dx.doi.org/10.1016/j.solener.2017.09.057. URL <http://www.sciencedirect.com/science/article/pii/S0038092X17308447> .
Janjai, S., Sricharoen, K., Pattarapanitchai, S., 2011. Semi-empirical models for the estimation of clear sky solar global and direct normal irradiances in the tropics. Appl. Energy 88, 4749–4755. http://dx.doi.org/10.1016/j.apenergy.2011.06.021. URL <http://www.sciencedirect.com/science/article/pii/S0306261911004090> .
Jiang, H., Dong, Y., 2016. A nonlinear support vector machine model with hard penalty function based on glowworm swarm optimization for forecasting daily global solar radiation. Energy Convers. Manage. 126, 991–1002. http://dx.doi.org/10.1016/j.enconman.2016.08.069. URL <http://www.sciencedirect.com/science/article/pii/S0196890416307439> .
Jiang, H., Dong, Y., 2017. Forecast of hourly global horizontal irradiance based on structured kernel support vector machine: A case study of Tibet area in China. Energy Convers. Manage. 142, 307–321. http://dx.doi.org/10.1016/j.enconman.2017.03.054. URL <http://www.sciencedirect.com/science/article/pii/S0196890417302650> .
Jiang, H., Dong, Y., Wang, J., Li, Y., 2015. Intelligent optimization models based on hard-ridge penalty and RBF for forecasting global solar radiation. Energy Convers. Manage. 95, 42–58. http://dx.doi.org/10.1016/j.enconman.2015.02.020. URL <http://www.sciencedirect.com/science/article/pii/S0196890415001338> .
Jiang, H., Dong, Y., Xiao, L., 2017. A multi-stage intelligent approach based on an ensemble of two-way interaction model for forecasting the global horizontal radiation of india. Energy Convers. Manage. 137, 142–154. http://dx.doi.org/10.1016/j.enconman.2017.01.040. URL <http://www.sciencedirect.com/science/article/pii/S0196890417300481> .
Jimenez, P.A., Hacker, J.P., Dudhia, J., Haupt, S.E., Ruiz-Arias, J.A., Gueymard, C.A., Thompson, G., Eidhammer, T., Deng, A., 2016. WRF-Solar: Description and clear-sky assessment of an augmented NWP model for solar power prediction. Bull. Am. Meteorol. Soc. 97, 1249–1264. http://dx.doi.org/10.1175/BAMS-D-14-00279.1.
Jolliffe, I.T., 1986. Principal component analysis and factor analysis. In: Principal Component Analysis. Springer, pp. 115–128.
Karakaya, E., 2016. Finite element method for forecasting the diffusion of photovoltaic systems: Why and how? Appl. Energy 163, 464–475. http://dx.doi.org/10.1016/j.apenergy.2015.10.188. URL <http://www.sciencedirect.com/science/article/pii/S0306261915014403> .
Kariya, T., Kurata, H., 2004. Generalized Least Squares. Wiley.
Kashyap, Y., Bansal, A., Sao, A.K., 2015. Solar radiation forecasting with multiple parameters neural networks. Renew. Sustain. Energy Rev. 49, 825–835. http://dx.doi.org/10.1016/j.rser.2015.04.077. URL <http://www.sciencedirect.com/science/article/pii/S1364032115003470> .
Kausar, M.A., Dhaka, V.S., Singh, S.K., 2013. Web crawler: A review. Int. J. Comput. Appl. 63, 31–36.
Killinger, S., Braam, F., Müller, B., Wille-Haussmann, B., McKenna, R., 2016. Projection of power generation between differently-oriented PV systems. Solar Energy 136, 153–165. http://dx.doi.org/10.1016/j.solener.2016.06.075. URL <http://www.sciencedirect.com/science/article/pii/S0038092X16302559> .
Killinger, S., Engerer, N., Müller, B., 2017. QCPV: A quality control algorithm for distributed photovoltaic array power output. Solar Energy 143, 120–131. http://dx.doi.org/10.1016/j.solener.2016.12.053. URL <http://www.sciencedirect.com/science/article/pii/S0038092X16306600> .
Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P., 1983. Optimization by simulated annealing. science 220, 671–680.
Kleissl, J., 2013. Solar Energy Forecasting and Resource Assessment. Academic Press.
Krause, J., 2017. Introducing Regular Expressions. Apress.
Kuhn, P., Wilbert, S., Prahl, C., Schüler, D., Haase, T., Hirsch, T., Wittmann, M., Ramirez, L., Zarzalejo, L., Meyer, A., Vuilleumier, L., Blanc, P., Pitz-Paal, R., 2017. Shadow camera system for the generation of solar irradiance maps. Solar Energy 157, 157–170. http://dx.doi.org/10.1016/j.solener.2017.05.074. URL <http://www.sciencedirect.com/science/article/pii/S0038092X17304814> .
Kurtz, B., Kleissl, J., 2017. Measuring diffuse, direct, and global irradiance using a sky imager. Solar Energy 141, 311–322. http://dx.doi.org/10.1016/j.solener.2016.11.032. URL <http://www.sciencedirect.com/science/article/pii/S0038092X16305722> .
Kurtz, B., Mejia, F., Kleissl, J., 2017. A virtual sky imager testbed for solar energy forecasting. Solar Energy 158, 753–759. http://dx.doi.org/10.1016/j.solener.2017.10.036. URL <https://www.sciencedirect.com/science/article/pii/S0038092X1730899X> .
Kwartler, T., 2017. Text Mining in Practice with R. Wiley.
Kwiatkowski, D., Phillips, P.C.B., Schmidt, P., Shin, Y., 1992. Testing the null hypothesis of stationarity against the alternative of a unit root. J. Economet. 54, 159–178. http://dx.doi.org/10.1016/0304-4076(92)90104-Y. URL <http://www.sciencedirect.com/science/article/pii/030440769290104Y> .
Lara-Fanego, V., Ruiz-Arias, J.A., Pozo-Vázquez, D., Santos-Alamillos, F.J., Tovar-Pescador, J., 2012. Evaluation of the WRF model solar irradiance forecasts in Andalusia (southern Spain). Solar Energy 86, 2200–2217. http://dx.doi.org/10.1016/j.solener.2011.02.014. URL, progress in Solar Energy 3 <http://www.sciencedirect.com/science/article/pii/S0038092X11000582> .
Larson, D.P., Nonnenmacher, L., Coimbra, C.F.M., 2016. Day-ahead forecasting of solar power output from photovoltaic plants in the American Southwest. Renew. Energy 91, 11–20. http://dx.doi.org/10.1016/j.renene.2016.01.039. URL <http://www.sciencedirect.com/science/article/pii/S0960148116300398> .
Larson, V.E., 2013. Forecasting solar irradiance with numerical weather prediction models. In: Kleissl, J. (Ed.), Solar Energy Forecasting and Resource Assessment. Academic Press, pp. 299–318 (chapter 12).
Lauret, P., Voyant, C., Soubdhan, T., David, M., Poggi, P., 2015. A benchmarking of machine learning techniques for solar radiation forecasting in an insular context. Solar Energy 112, 446–457. http://dx.doi.org/10.1016/j.solener.2014.12.014. URL <http://www.sciencedirect.com/science/article/pii/S0038092X14006057> .
Lave, M., Kleissl, J., Stein, J.S., 2013. A wavelet-based variability model (WVM) for solar PV power plants. IEEE Trans. Sustain. Energy 4, 501–509. http://dx.doi.org/10.1109/TSTE.2012.2205716.
Law, E.W., Kay, M., Taylor, R.A., 2016a. Calculating the financial value of a concentrated solar thermal plant operated using direct normal irradiance forecasts. Solar Energy 125, 267–281. http://dx.doi.org/10.1016/j.solener.2015.12.031. URL <http://www.sciencedirect.com/science/article/pii/S0038092X15007045> .
Law, E.W., Kay, M., Taylor, R.A., 2016b. Evaluating the benefits of using short-term direct normal irradiance forecasts to operate a concentrated solar thermal plant. Solar Energy 140, 93–108. http://dx.doi.org/10.1016/j.solener.2016.10.037. URL <http://www.sciencedirect.com/science/article/pii/S0038092X16305023> .
Law, E.W., Prasad, A.A., Kay, M., Taylor, R.A., 2014. Direct normal irradiance forecasting and its application to concentrated solar thermal output forecasting – a review. Solar Energy 108, 287–307. http://dx.doi.org/10.1016/j.solener.2014.07.008. URL <http://www.sciencedirect.com/science/article/pii/S0038092X14003466> .
Lawrence, M.G., 2005. The relationship between relative humidity and the dewpoint temperature in moist air: A simple conversion and applications. Bulletin of the American Meteorological Society 86, 225–233. http://dx.doi.org/10.1175/BAMS-86-2-225.
Lee, G.M., Tam, N.N., Yen, N.D., 2005. Quadratic Programming and Affine Variational Inequalities. Springer.
Lefèvre, M., Oumbe, A., Blanc, P., Espinar, B., Gschwind, B., Qu, Z., Wald, L., Schroedter Homscheidt, M., Hoyer-Klick, C., Arola, A., Benedetti, A., Kaiser, J.W., Morcrette, J.J., 2013. McClear: A new model estimating downwelling solar radiation at ground level in clear-sky conditions. Atmos. Measur. Techniq. 6, 2403–2418. http://dx.doi.org/10.5194/amt-6-2403-2013. URL <https://hal-mines-paristech.archives-ouvertes.fr/hal-00862906> .
Li, C.H., Lee, C.K., 1993. Minimum cross entropy thresholding. Pattern Recog. 26, 617–625. http://dx.doi.org/10.1016/0031-3203(93)90115-D. URL <http://www.sciencedirect.com/science/article/pii/003132039390115D> .
Li, J., Ward, J.K., Tong, J., Collins, L., Platt, G., 2016a. Machine learning for solar irradiance forecasting of photovoltaic system. Renew. Energy 90, 542–553. http://dx.doi.org/10.1016/j.renene.2015.12.069. URL <http://www.sciencedirect.com/science/article/pii/S0960148115305747> .
Li, M., Chu, Y., Pedro, H.T.C., Coimbra, C.F.M., 2016b. Quantitative evaluation of the impact of cloud transmittance and cloud velocity on the accuracy of short-term DNI forecasts. Renew. Energy 86, 1362–1371. http://dx.doi.org/10.1016/j.renene.2015.09.058. URL <http://www.sciencedirect.com/science/article/pii/S0960148115303372> .
Li, Q., Lu, W., Yang, J., 2011. A hybrid thresholding algorithm for cloud detection on ground-based color images. J. Atmos. Ocean. Technol. 28, 1286–1296. http://dx.doi.org/10.1175/JTECH-D-11-00009.1.
Li, R., Zeng, B., Liou, M.L., 1994. A new three-step search algorithm for block motion estimation. IEEE Trans. Circ. Syst. Video Technol. 4, 438–442. http://dx.doi.org/10.1109/76.313138.
Li, Y., He, Y., Su, Y., Shu, L., 2016c. Forecasting the daily power output of a grid-connected photovoltaic system based on multivariate adaptive regression splines. Appl. Energy 180, 392–401. http://dx.doi.org/10.1016/j.apenergy.2016.07.052. URL <http://www.sciencedirect.com/science/article/pii/S0306261916309941> .
Li, Y., Su, Y., Shu, L., 2014. An ARMAX model for forecasting the power output of a grid connected photovoltaic system. Renew. Energy 66, 78–89. http://dx.doi.org/10.1016/j.renene.2013.11.067. URL <http://www.sciencedirect.com/science/article/pii/S0960148113006551> .
Lima, F.J.L., Martins, F.R., Pereira, E.B., Lorenz, E., Heinemann, D., 2016. Forecast for surface solar irradiance at the Brazilian Northeastern region using NWP model and artificial neural networks. Renew. Energy 87 (Part 1), 807–818. http://dx.doi.org/10.1016/j.renene.2015.11.005. URL <http://www.sciencedirect.com/science/article/pii/S0960148115304249> .
Lin, K.P., Pai, P.F., 2016. Solar power output forecasting using evolutionary seasonal decomposition least-square support vector regression. J. Clean. Product. 134 (Part B), 456–462. http://dx.doi.org/10.1016/j.jclepro.2015.08.099. URL, special Volume: Green and Sustainable Innovation for Cleaner Production in the Asia-Pacific Region <http://www.sciencedirect.com/science/article/pii/S0959652615012007> .
Liu, Y., Shimada, S., Yoshino, J., Kobayashi, T., Miwa, Y., Furuta, K., 2016. Ensemble forecasting of solar irradiance by applying a mesoscale meteorological model. Solar

Energy 136, 597–605. http://dx.doi.org/10.1016/j.solener.2016.07.043. URL <http://www.sciencedirect.com/science/article/pii/S0038092X16303024> .
Long, C.N., Slater, D.W., Tooman, T., 2001. Total sky imager model 880 status and testing results. Technical Report. DOE Office of Science Atmospheric Radiation Measurement (ARM) Program (United States).
Lonij, V.P.A., Brooks, A.E., Cronin, A.D., Leuthold, M., Koch, K., 2013. Intra-hour forecasts of solar power production using measurements from a network of irradiance sensors. Solar Energy 97, 58–66. http://dx.doi.org/10.1016/j.solener.2013.08.002. URL <http://www.sciencedirect.com/science/article/pii/S0038092X13003125> .
López-Cózar, E.D., Robinson-García, N., Torres-Salinas, D., 2014. The Google Scholar experiment: How to index false papers and manipulate bibliometric indicators. J. Assoc. Inf. Sci. Technol. 65, 446–454. http://dx.doi.org/10.1002/asi.23056.
Lorenzo, A.T., Holmgren, W.F., Cronin, A.D., 2015. Irradiance forecasts based on an irradiance monitoring network, cloud motion, and spatial averaging. Solar Energy 122, 1158–1169. http://dx.doi.org/10.1016/j.solener.2015.10.038. URL <http://www.sciencedirect.com/science/article/pii/S0038092X15005897> .
Lorenzo, A.T., Morzfeld, M., Holmgren, W.F., Cronin, A.D., 2017. Optimal interpolation of satellite and ground data for irradiance nowcasting at city scales. Solar Energy 144, 466–474. http://dx.doi.org/10.1016/j.solener.2017.01.038. URL <http://www.sciencedirect.com/science/article/pii/S0038092X17300555> .
Lujano-Rojas, J.M., Osório, G.J., Matias, J.C.O., Catalão, J.P.S., 2016. A heuristic methodology to economic dispatch problem incorporating renewable power forecasting error and system reliability. Renew. Energy 87 (Part 1), 731–743. http://dx.doi.org/10.1016/j.renene.2015.11.011. URL <http://www.sciencedirect.com/science/article/pii/S0960148115304316> .
Mahoney, W.P., Parks, K., Wiener, G., Liu, Y., Myers, W.L., Sun, J., Monache, L.D., Hopson, T., Johnson, D., Haupt, S.E., 2012. A wind power forecasting system to optimize grid integration. IEEE Trans. Sustain. Energy 3, 670–682. http://dx.doi.org/10.1109/TSTE.2012.2201758.
Manobianco, J., Taylor, G.E., Zack, J.W., 1996. Workstation-based real-time mesoscale modeling designed for weather support to operations at the Kennedy Space Center and Cape Canaveral Air Station. Bull. Am. Meteorol. Soc. 77, 653–672. http://dx.doi.org/10.1175/1520-0477(1996)077<0653:WBRTMM>2.0.CO;2.
Marion, B., 2015. A model for deriving the direct normal and diffuse horizontal irradiance from the global tilted irradiance. Solar Energy 122, 1037–1046. http://dx.doi.org/10.1016/j.solener.2015.10.024. URL <http://www.sciencedirect.com/science/article/pii/S0038092X15005757> .
Marquez, R., Coimbra, C.F.M., 2011. Forecasting of global and direct solar irradiance using stochastic learning methods, ground experiments and the NWS database. Solar Energy 85, 746–756. http://dx.doi.org/10.1016/j.solener.2011.01.007. URL <http://www.sciencedirect.com/science/article/pii/S0038092X11000193> .
Marquez, R., Coimbra, C.F.M., 2012. Proposed metric for evaluation of solar forecasting models. J. Solar Energy Eng. 135, 011016–011016 – 9. http://dx.doi.org/10.1115/1.4007496.
Marquez, R., Coimbra, C.F.M., 2013. Intra-hour DNI forecasting based on cloud tracking image analysis. Solar Energy 91, 327–336. http://dx.doi.org/10.1016/j.solener.2012.09.018. URL <http://www.sciencedirect.com/science/article/pii/S0038092X1200343X> .
Marquez, R., Pedro, H.T.C., Coimbra, C.F.M., 2013. Hybrid solar forecasting method uses satellite imaging and ground telemetry as inputs to ANNs. Solar Energy 92, 176–188. http://dx.doi.org/10.1016/j.solener.2013.02.023. URL <http://www.sciencedirect.com/science/article/pii/S0038092X13000881> .
Marvin, C., Kimball, H., 1926. Solar radiation and weather forecasting. J. Frank. Inst. 202, 273–306. http://dx.doi.org/10.1016/S0016-0032(26)91369-0. URL <http://www.sciencedirect.com/science/article/pii/S0016003226913690> .
Massey, F.J., 1951. The Kolmogorov–Smirnov test for goodness of fit. J. Am. Statist. Assoc. 46, 68–78. http://dx.doi.org/10.2307/2280095. URL <http://www.jstor.org/stable/2280095> .
Massidda, L., Marrocu, M., 2017. Use of multilinear adaptive regression splines and numerical weather prediction to forecast the power output of a PV plant in Borkum, Germany. Solar Energy 146, 141–149. http://dx.doi.org/10.1016/j.solener.2017.02.007. URL <http://www.sciencedirect.com/science/article/pii/S0038092X17301007> .
Matheson, J.E., Winkler, R.L., 1976. Scoring rules for continuous probability distributions. Manage. Sci. 22, 1087–1096. http://dx.doi.org/10.1287/mnsc.22.10.1087.
Mathiesen, P., Kleissl, J., 2011. Evaluation of numerical weather prediction for intra-day solar forecasting in the continental United States. Solar Energy 85, 967–977. http://dx.doi.org/10.1016/j.solener.2011.02.013. URL <http://www.sciencedirect.com/science/article/pii/S0038092X11000570> .
Mazorra Aguiar, L., Pereira, B., Lauret, P., Díaz, F., David, M., 2016. Combining solar irradiance measurements, satellite-derived data and a numerical weather prediction model to improve intra-day solar forecasting. Renew. Energy 97, 599–610. http://dx.doi.org/10.1016/j.renene.2016.06.018. URL <http://www.sciencedirect.com/science/article/pii/S0960148116305390> .
McCandless, T.C., Haupt, S.E., Young, G.S., 2016. A regime-dependent artificial neural network technique for short-range solar irradiance forecasting. Renew. Energy 89, 351–359. http://dx.doi.org/10.1016/j.renene.2015.12.030. URL <http://www.sciencedirect.com/science/article/pii/S0960148115305346> .
McCulloch, W.S., Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. Bull. Math. Biophys. 5, 115–133.
Mejia, F.A., Kurtz, B., Kleissl, J., 2018. Cloud tomography applied to sky images: Part I: A virtual testbed (in preparation).
Mejia, F.A., Kurtz, B., Murray, K., Hinkelman, L.M., Sengupta, M., Xie, Y., Kleissl, J., 2015. Coupling sky images with three-dimensional radiative transfer models: A new method to estimate cloud optical depth. Atmos. Measur. Techn. Disc. 8, 11285–11321.
Mellit, A., Benghanem, M., Kalogirou, S.A., 2006. An adaptive wavelet-network model for forecasting daily total solar-radiation. Appl. Energy 83, 705–722. http://dx.doi.org/10.1016/j.apenergy.2005.06.003. URL <http://www.sciencedirect.com/science/article/pii/S0306261905000875> .
Mellit, A., Pavan, A.M., 2010. A 24-h forecast of solar irradiance using artificial neural network: Application for performance prediction of a grid-connected PV plant at Trieste, Italy. Solar Energy 84, 807–821. http://dx.doi.org/10.1016/j.solener.2010.02.006. URL <http://www.sciencedirect.com/science/article/pii/S0038092X10000782> .
Mellit, A., Pavan, A.M., Lughi, V., 2014. Short-term forecasting of power production in a large-scale photovoltaic plant. Solar Energy 105, 401–413. http://dx.doi.org/10.1016/j.solener.2014.03.018. URL <http://www.sciencedirect.com/science/article/pii/S0038092X14001522> .
Miles, J.W., 1971. Integral Transforms in Applied Mathematics. Cambridge University Press.
Miller, S.D., Rogers, M.A., Haynes, J.M., Sengupta, M., Heidinger, A.K., 2018. Short-term solar irradiance forecasting via satellite/model coupling. Solar Energy. http://dx.doi.org/10.1016/j.solener.2017.11.049. URL <https://www.sciencedirect.com/science/article/pii/S0038092X17310435> .
Minsky, M., Papert, S., 1969. Perceptrons. MIT Press.
Monjoly, S., André, M., Calif, R., Soubdhan, T., 2017. Hourly forecasting of global solar radiation based on multiscale decomposition methods: A hybrid approach. Energy 119, 288–298. http://dx.doi.org/10.1016/j.energy.2016.11.061. URL <http://www.sciencedirect.com/science/article/pii/S0360544216316668> .
Mora-López, L., Martínez-Marchena, I., Piliougine, M., Sidrach-deCardona, M., 2011. Machine learning approach for next day energy production forecasting in grid connected photovoltaic plants. In: World Renewable Energy Congress - Sweden; 8-13 May; 2011; Linköping; Sweden, Linköping University Electronic Press; Linköpings universitet, pp. 2869–2874.
Morcrette, J.J., Barker, H.W., Cole, J.N.S., Iacono, M.J., Pincus, R., 2008. Impact of a new radiation package, McRad, in the ECMWF integrated forecasting system. Mont. Weath. Rev. 136, 4773–4798. http://dx.doi.org/10.1175/2008MWR2363.1.
Nakamura, J., 2016. Image Sensors and Signal Processing for Digital Still Cameras. CRC Press.
Nguyen, D.A., Kleissl, J., 2014. Stereographic methods for cloud base height determination using two sky imagers. Solar Energy 107, 495–509. http://dx.doi.org/10.1016/j.solener.2014.05.005. URL <http://www.sciencedirect.com/science/article/pii/S0038092X14002333> .
Ni, Q., Zhuang, S., Sheng, H., Kang, G., Xiao, J., 2017. An ensemble prediction intervals approach for short-term PV power forecasting. Solar Energy 155, 1072–1083. http://dx.doi.org/10.1016/j.solener.2017.07.052. URL <http://www.sciencedirect.com/science/article/pii/S0038092X17306370> .
Nobre, A.M., Severiano Jr., C.A., Karthik, S., Kubis, M., Zhao, L., Martins, F.R., Pereira, E.B., Rüther, R., Reindl, T., 2016. PV power conversion and short-term forecasting in a tropical, densely-built environment in Singapore. Renew. Energy 94, 496–509. http://dx.doi.org/10.1016/j.renene.2016.03.075. URL <http://www.sciencedirect.com/science/article/pii/S0960148116302592> .
Nocedal, J., Wright, S.J., 1999. Numerical Optimization. Springer, New York.
Nonnenmacher, L., Coimbra, C.F.M., 2014. Streamline-based method for intra-day solar forecasting through remote sensing. Solar Energy 108, 447–459. http://dx.doi.org/10.1016/j.solener.2014.07.026. URL <http://www.sciencedirect.com/science/article/pii/S0038092X14003752> .
Nonnenmacher, L., Kaur, A., Coimbra, C.F.M., 2016. Day-ahead resource forecasting for concentrated solar power integration. Renew. Energy 86, 866–876. http://dx.doi.org/10.1016/j.renene.2015.08.068. URL <http://www.sciencedirect.com/science/article/pii/S0960148115302688> .
Nova, J.C., Cunha, J.B., de Moura Oliveira, P.B., 2005. Solar irradiation forecast model using time series analysis and sky images. In: Proceedings of the 5th Conference of the European Federation for Information Technology in Agriculture, Food and Environment, pp. 1408–1415.
Ogliari, E., Dolara, A., Manzolini, G., Leva, S., 2017. Physical and hybrid methods comparison for the day ahead PV output power forecast. Renew. Energy 113, 11–21. http://dx.doi.org/10.1016/j.renene.2017.05.063. URL <http://www.sciencedirect.com/science/article/pii/S096014811730455X> .
Ohtake, H., da Silva Fonseca, J.G., Takashima, T., Oozeki, T., ichi Shimose, K., Yamada, Y., 2015. Regional and seasonal characteristics of global horizontal irradiance forecasts obtained from the Japan Meteorological Agency mesoscale model. Solar Energy 116, 83–99. http://dx.doi.org/10.1016/j.solener.2015.03.020. URL <http://www.sciencedirect.com/science/article/pii/S0038092X15001383> .
Ohtake, H., ichi Shimose, K., da Silva Fonseca Jr., J.G., Takashima, T., Oozeki, T., Yamada, Y., 2013. Accuracy of the solar irradiance forecasts of the Japan Meteorological Agency mesoscale model for the Kanto region, Japan. Solar Energy 98 (Part B), 138–152. http://dx.doi.org/10.1016/j.solener.2012.10.007. URL, {ICEM} Solar Radiation <http://www.sciencedirect.com/science/article/pii/S0038092X12003611> .
Olden, J.D., Joy, M.K., Death, R.G., 2004. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. Ecol. Model. 178, 389–397. http://dx.doi.org/10.1016/j.ecolmodel.2004.03.013. URL <http://www.sciencedirect.com/science/article/pii/S0304380004001565> .
Ortega, J.L., Aguillo, I.F., 2014. Microsoft Academic Search and Google Scholar citations: Comparative analysis of author profiles. J. Assoc. Inf. Sci. Technol. 65, 1149–1156. http://dx.doi.org/10.1002/asi.23036.
Paulescu, M., Brabec, M., Boata, R., Badescu, V., 2017. Structured, physically inspired (gray box) models versus black box modeling for forecasting the output power of photovoltaic plants. Energy 121, 792–802. http://dx.doi.org/10.1016/j.energy.2017.01.015. URL <http://www.sciencedirect.com/science/article/pii/

S0360544217300154> .
Paulescu, M., Paulescu, E., Gravila, P., Badescu, V., 2013. Solar radiation measurements. In: Weather Modeling and Forecasting of PV Systems Operation. Springer, pp. 17–42.
Pecenak, Z.K., Mejia, F.A., Kurtz, B., Evan, A., Kleissl, J., 2016. Simulating irradiance enhancement dependence on cloud optical depth and solar zenith angle. Solar Energy 136, 675–681. http://dx.doi.org/10.1016/j.solener.2016.07.045. URL <http://www.sciencedirect.com/science/article/pii/S0038092X16303036> .
Pedro, H.T.C., Coimbra, C.F.M., 2012. Assessment of forecasting techniques for solar power production with no exogenous inputs. Solar Energy 86, 2017–2028. http://dx.doi.org/10.1016/j.solener.2012.04.004. URL <http://www.sciencedirect.com/science/article/pii/S0038092X12001429> .
Pelland, S., Galanis, G., Kallos, G., 2013. Solar and photovoltaic forecasting through post-processing of the global environmental multiscale numerical weather prediction model. Prog. Photovolt.: Res. Appl. 21, 284–296. http://dx.doi.org/10.1002/pip.1180.
Peng, Z., Yu, D., Huang, D., Heiser, J., Yoo, S., Kalb, P., 2015. 3D cloud detection and tracking system for solar forecast using multiple sky imagers. Solar Energy 118, 496–519. http://dx.doi.org/10.1016/j.solener.2015.05.037. URL <http://www.sciencedirect.com/science/article/pii/S0038092X15002972> .
Perez, M.J., Fthenakis, V.M., 2015. On the spatial decorrelation of stochastic solar resource variability at long timescales. Solar Energy 117, 46–58. http://dx.doi.org/10.1016/j.solener.2015.04.020. URL <http://www.sciencedirect.com/science/article/pii/S0038092X1500208X> .
Perez, R., David, M., Hoff, T.E., Jamaly, M., Kivalov, S., Kleissl, J., Lauret, P., Perez, M., et al., 2016. Spatial and temporal variability of solar energy. Found. Trends® Renew. Energy 1, 1–44.
Perez, R., Ineichen, P., Moore, K., Kmiecik, M., Chain, C., George, R., Vignola, F., 2002. A new operational model for satellite-derived irradiances: Description and validation. Solar Energy 73, 307–317. http://dx.doi.org/10.1016/S0038-092X(02)00122-6. URL <http://www.sciencedirect.com/science/article/pii/S0038092X02001226> .
Perez, R., Ineichen, P., Seals, R., Zelenka, A., 1990. Making full use of the clearness index for parameterizing hourly insolation conditions. Solar Energy 45, 111–114. http://dx.doi.org/10.1016/0038-092X(90)90036-C. URL <http://www.sciencedirect.com/science/article/pii/0038092X9090036C> .
Perez, R., Kivalov, S., Schlemmer, J., Hemker Jr., K., Renné, D., Hoff, T.E., 2010. Validation of short and medium term operational solar radiation forecasts in the US. Solar Energy 84, 2161–2172. http://dx.doi.org/10.1016/j.solener.2010.08.014. URL <http://www.sciencedirect.com/science/article/pii/S0038092X10002823> .
Perez, R., Lorenz, E., Pelland, S., Beauharnois, M., Knowe, G.V., Hemker Jr., K., Heinemann, D., Remund, J., Müller, S.C., Traunmüller, W., Steinmauer, G., Pozo, D., Ruiz-Arias, J.A., Lara-Fanego, V., Ramirez-Santigosa, L., Gaston-Romero, M., Pomares, L.M., 2013. Comparison of numerical weather prediction solar irradiance forecasts in the US, Canada and Europe. Solar Energy 94, 305–326. http://dx.doi.org/10.1016/j.solener.2013.05.005. URL <http://www.sciencedirect.com/science/article/pii/S0038092X13001886> .
Persson, C., Bacher, P., Shiga, T., Madsen, H., 2017. Multi-site solar power forecasting using gradient boosted regression trees. Solar Energy 150, 423–436. http://dx.doi.org/10.1016/j.solener.2017.04.066. URL <http://www.sciencedirect.com/science/article/pii/S0038092X17303717> .
Pfister, G., McKenzie, R.L., Liley, J.B., Thomas, A., Forgan, B.W., Long, C.N., 2003. Cloud coverage based on all-sky imaging and its impact on surface solar irradiance. J. Appl. Meteorol. 42, 1421–1434. http://dx.doi.org/10.1175/1520-0450(2003)042<1421:CCBOAI>2.0.CO;2.
Pielke, R.A., Cotton, W.R., Walko, R.L., Tremback, C.J., Lyons, W.A., Grasso, L.D., Nicholls, M.E., Moran, M.D., Wesley, D.A., Lee, T.J., Copeland, J.H., 1992. A comprehensive meteorological modeling system–RAMS. Meteorol. Atmos. Phys. 49, 69–91. http://dx.doi.org/10.1007/BF01025401.
Pierro, M., Bucci, F., Felice, M.D., Maggioni, E., Moser, D., Perotto, A., Spada, F., Cornaro, C., 2016. Multi-model ensemble for day ahead prediction of photovoltaic power generation. Solar Energy 134, 132–146. http://dx.doi.org/10.1016/j.solener.2016.04.040. URL <http://www.sciencedirect.com/science/article/pii/S0038092X16300731> .
Polo, J., Wilbert, S., Ruiz-Arias, J., Meyer, R., Gueymard, C., Súri, M., Martín, L., Mieslinger, T., Blanc, P., Grant, I., Boland, J., Ineichen, P., Remund, J., Escobar, R., Troccoli, A., Sengupta, M., Nielsen, K., Renné, D., Geuder, N., Cebecauer, T., 2016. Preliminary survey on site-adaptation techniques for satellite-derived and reanalysis solar radiation datasets. Solar Energy 132, 25–37. http://dx.doi.org/10.1016/j.solener.2016.03.001. URL <http://www.sciencedirect.com/science/article/pii/S0038092X16001754> .
Polo, J., Zarzalejo, L.F., Ramírez, L., 2008. Solar radiation derived from satellite images. In: Badescu, V. (Ed.), Modeling Solar Radiation at the Earth's Surface: Recent Advances. Springer, Berlin, Heidelberg, pp. 449–462. http://dx.doi.org/10.1007/978-3-540-77455-6_18.
Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of population structure using multilocus genotype data. Genetics 155, 945–959.
Qazi, A., Fayaz, H., Wadi, A., Raj, R.G., Rahim, N.A., Khan, W.A., 2015. The artificial neural network for solar radiation prediction and designing solar systems: A systematic literature review. J. Clean. Product. 104, 1–12. http://dx.doi.org/10.1016/j.jclepro.2015.04.041. URL <http://www.sciencedirect.com/science/article/pii/S0959652615004096> .
Qian, P.Z.G., Wu, C.F.J., 2008. Bayesian hierarchical modeling for integrating low-accuracy and high-accuracy experiments. Technometrics 50, 192–204. http://dx.doi.org/10.1198/004017008000000082.
Qu, Z., Oumbe, A., Blanc, P., Espinar, B., Gesell, G., Gschwind, B., Klüser, L., Lefèvre, M., Saboret, L., Schroedter-Homscheidt, M., Wald, L., 2017. Fast radiative transfer parameterisation for assessing the surface solar irradiance: The Heliosat-4 method. Meteorologische Zeitschrift 26, 33–57. http://dx.doi.org/10.1127/metz/2016/0781. URL <https://hal.archives-ouvertes.fr/hal-01512589> .
Quan, H., Srinivasan, D., Khosravi, A., 2014. Short-term load and wind power forecasting using neural network-based prediction intervals. IEEE Trans. Neural Netw. Learn. Syst. 25, 303–315. http://dx.doi.org/10.1109/TNNLS.2013.2276053.
Quesenberry, C.P., 2006. Probability integral transformations. In: Kotz, S., Read, C.B., Balakrishnan, N., Vidakovic, B. (Eds.), Encyclopedia of Statistical Sciences. John Wiley & Sons Inc.. http://dx.doi.org/10.1002/0471667196.ess2067.pub2.
Ramirez-Rosado, I.J., Fernandez-Jimenez, L.A., Monteiro, C., Garcia-Garrido, E., Zorzano-Santamaria, P., 2011. Spatial long-term forecasting of small power photovoltaic systems expansion. Renew. Energy 36, 3499–3506. http://dx.doi.org/10.1016/j.renene.2011.05.037. URL <http://www.sciencedirect.com/science/article/pii/S0960148111002904> .
Rana, M., Koprinska, I., Agelidis, V.G., 2015. 2D-interval forecasts for solar power production. Solar Energy 122, 191–203. http://dx.doi.org/10.1016/j.solener.2015.08.018. URL <http://www.sciencedirect.com/science/article/pii/S0038092X15004545> .
Rana, M., Koprinska, I., Agelidis, V.G., 2016. Univariate and multivariate methods for very short-term solar photovoltaic power forecasting. Energy Convers. Manage. 121, 380–390. http://dx.doi.org/10.1016/j.enconman.2016.05.025. URL <http://www.sciencedirect.com/science/article/pii/S0196890416303934> .
Rasmussen, C.E., Williams, C.K.I., 2006. Gaussian Processes for Machine Learning, vol. 1 MIT press, Cambridge.
Raza, M.Q., Nadarajah, M., Ekanayake, C., 2016. On recent advances in PV output power forecast. Solar Energy 136, 125–144. http://dx.doi.org/10.1016/j.solener.2016.06.073. URL <http://www.sciencedirect.com/science/article/pii/S0038092X16302547> .
Reda, I., Andreas, A., 2004. Solar position algorithm for solar radiation applications. Solar Energy 76, 577–589. http://dx.doi.org/10.1016/j.solener.2003.12.003. URL <http://www.sciencedirect.com/science/article/pii/S0038092X0300450X> .
Reikard, G., 2009. Predicting solar radiation at high resolutions: A comparison of time series forecasts. Solar Energy 83, 342–349. http://dx.doi.org/10.1016/j.solener.2008.08.007. URL <http://www.sciencedirect.com/science/article/pii/S0038092X08002107> .
Reikard, G., Haupt, S.E., Jensen, T., 2017. Forecasting ground-level irradiance over short horizons: Time series, meteorological, and time-varying parameter models. Renew. Energy 112, 474–485. http://dx.doi.org/10.1016/j.renene.2017.05.019. URL <http://www.sciencedirect.com/science/article/pii/S0960148117304044> .
Reinhard, E., Heidrich, W., Debevec, P., Pattanaik, S., Ward, G., Myszkowski, K., 2010. High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting. Morgan Kaufmann.
Ren, Y., Suganthan, P., Srikanth, N., 2015. Ensemble methods for wind and solar power forecasting–a state-of-the-art review. Renew. Sustain. Energy Rev. 50, 82–91. http://dx.doi.org/10.1016/j.rser.2015.04.081. URL <http://www.sciencedirect.com/science/article/pii/S1364032115003512> .
Rosenblatt, F., 1958. The perceptron: A probabilistic model for information storage and organization in the brain. Psychol. Rev. 65, 386–408.
Sahai, H., Ojeda, M.M., 2004. Analysis of variance for random models. Balanced Data Theory, Methods, Applications and Data Analysis, vol. I Springer.
Sakoe, H., Chiba, S., 1978. Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. Acoust., Speech, Sig. Process. 26, 43–49. http://dx.doi.org/10.1109/TASSP.1978.1163055.
Salcedo-Sanz, S., Casanova-Mateo, C., Pastor-Sánchez, A., Sánchez-Girón, M., 2014. Daily global solar radiation prediction based on a hybrid coral reefs optimization – extreme learning machine approach. Solar Energy 105, 91–98. http://dx.doi.org/10.1016/j.solener.2014.04.009. URL <http://www.sciencedirect.com/science/article/pii/S0038092X14001947> .
Salomaa, A., 1969. Chapter II – Finite non-deterministic and probabilistic automata. In: Salomaa, A. (Ed.), Theory of Automata. International Series of Monographs on Pure and Applied Mathematics, vol. 100. Pergamon, pp. 71–113. http://dx.doi.org/10.1016/B978-0-08-013376-8.50008-3. URL <http://www.sciencedirect.com/science/article/pii/B9780080133768500083> .
Sanfilippo, A., Martin-Pomares, L., Mohandes, N., Perez-Astudillo, D., Bachour, D., 2016. An adaptive multi-modeling approach to solar nowcasting. Solar Energy 125, 77–85. http://dx.doi.org/10.1016/j.solener.2015.11.041. URL <http://www.sciencedirect.com/science/article/pii/S0038092X15006581> .
Schalkwijk, J., Jonker, H.J.J., Siebesma, A.P., Meijgaard, E.V., 2015. Weather forecasting using GPU-based large-eddy simulations. Bull. Am. Meteorol. Soc. 96, 715–723. http://dx.doi.org/10.1175/BAMS-D-14-00114.1.
Scharmer, K., Greif, J., Dogniaux, R., 2000. The European Solar Radiation Atlas. Les Presses de l'École des Mines.
Schmidt, T., Kalisch, J., Lorenz, E., Heinemann, D., 2016. Evaluating the spatio-temporal performance of sky-imager-based solar irradiance analysis and forecasts. Atmos. Chem. Phys. 16, 3399–3412. http://dx.doi.org/10.5194/acp-16-3399-2016. URL <http://www.atmos-chem-phys.net/16/3399/2016/> .
Schwartz, A.S., Hearst, M.A., 2003. A simple algorithm for identifying abbreviation definitions in biomedical text. In: Pacific Symposium on Biocomputing. Kauai, Hawaii, pp. 451–462.
Sengupta, M., Habte, A., Gueymard, C., Wilbert, S., Renne, D., 2017. Best Practices Handbook for the Collection and Use of Solar Resource Data for Solar Energy Applications. Technical Report NREL/TP-5D00-68886. National Renewable Energy Lab < https://www.nrel.gov/docs/fy18osti/68886.pdf > .
Sepasi, S., Reihani, E., Howlader, A.M., Roose, L.R., Matsuura, M.M., 2017. Very short term load forecasting of a distribution system with high PV penetration. Renew. Energy 106, 142–148. http://dx.doi.org/10.1016/j.renene.2017.01.019. URL <http://www.sciencedirect.com/science/article/pii/S0960148117300198> .

Serenko, A., Dumay, J., 2015. Citation classics published in Knowledge Management journals. Part II: Studying research trends and discovering the Google Scholar effect. J. Knowl. Manage. 19, 1335–1355. http://dx.doi.org/10.1108/JKM-02-2015-0086.
Sfetsos, A., Coonick, A., 2000. Univariate and multivariate forecasting of hourly solar radiation with artificial intelligence techniques. Solar Energy 68, 169–178. http://dx.doi.org/10.1016/S0038-092X(99)00064-X. URL <http://www.sciencedirect.com/science/article/pii/S0038092X9900064X> .
Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27, 379–423.
Shariff, S.Z., Bejaimal, S.A.D., Sontrop, J.M., Iansavichus, A.V., Haynes, B., Weir, M.A., Garg, A.X., 2013. Retrieving clinical evidence: A comparison of PubMed and Google Scholar for quick clinical searches. J. Med. Internet Res. 15, e164. http://dx.doi.org/10.2196/jmir.2624.
Sharma, V., Yang, D., Walsh, W., Reindl, T., 2016. Short term solar irradiance forecasting using a mixed wavelet neural network. Renew. Energy 90, 481–492. http://dx.doi.org/10.1016/j.renene.2016.01.020. URL <http://www.sciencedirect.com/science/article/pii/S0960148116300209> .
Singh, A.V., Vikas, A.M., 2014. A review of web crawler algorithms. Int. J. Comp. Sci. Inf. Technol. 5, 6689–6691.
Soubdhan, T., Ndong, J., Ould-Baba, H., Do, M.T., 2016. A robust forecasting framework based on the Kalman filtering approach with a twofold parameter tuning procedure: Application to solar and photovoltaic prediction. Solar Energy 131, 246–259. http://dx.doi.org/10.1016/j.solener.2016.02.036. URL <http://www.sciencedirect.com/science/article/pii/S0038092X16001444> .
Sperati, S., Alessandrini, S., Monache, L.D., 2016. An application of the ECMWF ensemble prediction system for short-term solar power forecasting. Solar Energy 133, 437–450. http://dx.doi.org/10.1016/j.solener.2016.04.016. URL <http://www.sciencedirect.com/science/article/pii/S0038092X1630041X> .
Storn, R., Price, K., 1997. Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. J. Global Optim. 11, 341–359. http://dx.doi.org/10.1023/A:1008202821328.
Sun, H., Yan, D., Zhao, N., Zhou, J., 2015. Empirical investigation on modeling solar radiation series with ARMA–GARCH models. Energy Convers. Manage. 92, 385–395. http://dx.doi.org/10.1016/j.enconman.2014.12.072. URL <http://www.sciencedirect.com/science/article/pii/S019689041401111X> .
Suter II, G.W., 2013. Review papers are important and worth writing. Environ. Toxicol. Chem. 32, 1929–1930. http://dx.doi.org/10.1002/etc.2316.
Takeda, H., 2017. Short-term ensemble forecast for purchased photovoltaic generation. Solar Energy 149, 176–187. http://dx.doi.org/10.1016/j.solener.2017.03.088. URL <http://www.sciencedirect.com/science/article/pii/S0038092X17302785> .
Tapakis, R., Charalambides, A.G., 2013. Equipment and methodologies for cloud detection and classification: A review. Solar Energy 95, 392–430. http://dx.doi.org/10.1016/j.solener.2012.11.015. URL <http://www.sciencedirect.com/science/article/pii/S0038092X12004069> .
Thompson, K., 1968. Programming techniques: Regular expression search algorithm. Commun. ACM 11, 419–422. http://dx.doi.org/10.1145/363347.363387.
Thorey, J., Mallet, V., Chaussin, C., Descamps, L., Blanc, P., 2015. Ensemble forecast of solar radiation using TIGGE weather forecasts and HelioClim database. Solar Energy 120, 232–243. http://dx.doi.org/10.1016/j.solener.2015.06.049. URL <http://www.sciencedirect.com/science/article/pii/S0038092X15003576> .
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Statist. Soc. Ser. B (Methodol.) 58, 267–288. URL <http://www.jstor.org/stable/2346178> .
Tödter, J., Ahrens, B., 2012. Generalization of the ignorance score: Continuous ranked version and its decomposition. Mont. Weath. Rev. 140, 2005–2017. http://dx.doi.org/10.1175/MWR-D-11-00266.1.
Urquhart, B., Kurtz, B., Dahlin, E., Ghonima, M., Shields, J.E., Kleissl, J., 2015. Development of a sky imaging system for short-term solar power forecasting. Atmos. Measur. Techniq. 8, 875–890. http://dx.doi.org/10.5194/amt-8-875-2015. URL <https://www.atmos-meas-tech.net/8/875/2015/> .
Urraca, R., Antonanzas, J., Alia-Martinez, M., de Pison, F.J.M., Antonanzas-Torres, F., 2016. Smart baseline models for solar irradiation forecasting. Energy Convers. Manage. 108, 539–548. http://dx.doi.org/10.1016/j.enconman.2015.11.033. URL <http://www.sciencedirect.com/science/article/pii/S0196890415010535> .
Urraca, R., Gracia-Amillo, A.M., Huld, T., de Pison, F.J.M., Trentmann, J., Lindfors, A.V., Riihelä, A., Sanz-Garcia, A., 2017. Quality control of global solar radiation data with satellite-based products. Solar Energy 158, 49–62. http://dx.doi.org/10.1016/j.solener.2017.09.032. URL <http://www.sciencedirect.com/science/article/pii/S0038092X17308046> .
Vallance, L., Blanc, P., 2017. Personal communication.
Vallance, L., Charbonnier, B., Paul, N., Dubost, S., Blanc, P., 2017. Towards a standardized procedure to assess solar forecast accuracy: A new ramp and time alignment metric. Solar Energy 150, 408–422. http://dx.doi.org/10.1016/j.solener.2017.04.064. URL <http://www.sciencedirect.com/science/article/pii/S0038092X17303687> .
van der Meer, D.W., Widén, J., Munkhammar, J., 2017. Review on probabilistic forecasting of photovoltaic power production and electricity consumption. Renew. Sustain. Energy Rev. http://dx.doi.org/10.1016/j.rser.2017.05.212. URL <http://www.sciencedirect.com/science/article/pii/S1364032117308523> .
Verzijlbergh, R.A., Heijnen, P.W., de Roode, S.R., Los, A., Jonker, H.J.J., 2015. Improved model output statistics of numerical weather prediction based irradiance forecasts for solar power applications. Solar Energy 118, 634–645. http://dx.doi.org/10.1016/j.solener.2015.06.005. URL <http://www.sciencedirect.com/science/article/pii/S0038092X15003138> .
Vignola, F., Michalsky, J., Stoffel, T., 2012. Solar and Infrared Radiation Measurements. CRC Press.
Voyant, C., Motte, F., Fouilloy, A., Notton, G., Paoli, C., Nivet, M.L., 2017a. Forecasting method for global radiation time series without training phase: Comparison with other well-known prediction methodologies. Energy 120, 199–208. http://dx.doi.org/10.1016/j.energy.2016.12.118. URL <http://www.sciencedirect.com/science/article/pii/S0360544216319326> .
Voyant, C., Notton, G., Darras, C., Fouilloy, A., Motte, F., 2017b. Uncertainties in global radiation time series forecasting using machine learning: The multilayer perceptron case. Energy 125, 248–257. http://dx.doi.org/10.1016/j.energy.2017.02.098. URL <http://www.sciencedirect.com/science/article/pii/S0360544217302803> .
Voyant, C., Notton, G., Kalogirou, S., Nivet, M.L., Paoli, C., Motte, F., Fouilloy, A., 2017c. Machine learning methods for solar radiation forecasting: A review. Renew. Energy 105, 569–582. http://dx.doi.org/10.1016/j.renene.2016.12.095. URL <http://www.sciencedirect.com/science/article/pii/S0960148116311648> .
Wan, C., Zhao, J., Song, Y., Xu, Z., Lin, J., Hu, Z., 2015. Photovoltaic and solar power forecasting for smart grid energy management. CSEE J. Power Energy Syst. 1, 38–46. http://dx.doi.org/10.17775/CSEEJPES.2015.00046.
Wang, F., Zhen, Z., Mi, Z., Sun, H., Su, S., Yang, G., 2015. Solar irradiance feature extraction and support vector machines based weather status pattern recognition model for short-term photovoltaic power forecasting. Energy Build. 86, 427–438. http://dx.doi.org/10.1016/j.enbuild.2014.10.002. URL <http://www.sciencedirect.com/science/article/pii/S0378778814008226> .
Wang, G., Kurtz, B., Kleissl, J., 2016a. Cloud base height from sky imager and cloud speed sensor. Solar Energy 131, 208–221. http://dx.doi.org/10.1016/j.solener.2016.02.027. URL <http://www.sciencedirect.com/science/article/pii/S0038092X16001237> .
Wang, G., Su, Y., Shu, L., 2016b. One-day-ahead daily power forecasting of photovoltaic systems based on partial functional linear regression models. Renew. Energy 96 (Part A), 469–478. http://dx.doi.org/10.1016/j.renene.2016.04.089. URL <http://www.sciencedirect.com/science/article/pii/S0960148116303913> .
Wasserman, L., 2004. All of Statistics: A Concise Course in Statistical Inference. Springer.
Wasserman, L., 2006. All of Nonparametric Statistics. Springer.
Werbos, P.J., 1974. Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences, Ph.D. thesis. Harvard University.
Wilcox, S., Marion, W., 2008. Users Manual for TMY3 Data Sets. Technical Report NREL/TP-581-43156. National Renewable Energy Laboratory < https://www.nrel.gov/docs/fy08osti/43156.pdf > .
Wolff, B., Khnert, J., Lorenz, E., Kramer, O., Heinemann, D., 2016. Comparing support vector regression for PV power forecasting to a physical modeling approach using measurement, numerical weather prediction, and cloud motion data. Solar Energy 135, 197–208. http://dx.doi.org/10.1016/j.solener.2016.05.051. URL <http://www.sciencedirect.com/science/article/pii/S0038092X16301682> .
Wolsey, L.A., 2007. Mixed integer programming. In: Wiley Encyclopedia of Computer Science and Engineering. John Wiley & Sons Inc.. http://dx.doi.org/10.1002/9780470050118.ecse244.
Wu, J., Chee, K.C., 2011. Prediction of hourly solar radiation using a novel hybrid model of ARMA and TDNN. Solar Energy 85, 808–817. http://dx.doi.org/10.1016/j.solener.2011.01.013. URL <http://www.sciencedirect.com/science/article/pii/S0038092X11000259> .
Wu, X., Zhu, X., Wu, G.Q., Ding, W., 2014. Data mining with big data. IEEE Trans. Knowl. Data Eng. 26, 97–107. http://dx.doi.org/10.1109/TKDE.2013.109.
Xiong, S., Qian, P.Z.G., Wu, C.F.J., 2013. Sequential design and analysis of high-accuracy and low-accuracy computer codes. Technometrics 55, 37–46. http://dx.doi.org/10.1080/00401706.2012.723572.
Yang, D., 2016. Solar radiation on inclined surfaces: Corrections and benchmarks. Solar Energy 136, 288–302. http://dx.doi.org/10.1016/j.solener.2016.06.062. URL <http://www.sciencedirect.com/science/article/pii/S0038092X16302432> .
Yang, D., 2017. On adding and removing sensors in a solar irradiance monitoring network for areal forecasting and PV system performance evaluation. Solar Energy 155, 1417–1430. http://dx.doi.org/10.1016/j.solener.2017.07.061. URL <http://www.sciencedirect.com/science/article/pii/S0038092X17306461> .
Yang, D., Dong, Z., Lim, L.H.I., Liu, L., 2017a. Analyzing big time series data in solar engineering using features and PCA. Solar Energy 153, 317–328. http://dx.doi.org/10.1016/j.solener.2017.05.072. URL <http://www.sciencedirect.com/science/article/pii/S0038092X17304796> .
Yang, D., Dong, Z., Nobre, A., Khoo, Y.S., Jirutitijaroen, P., Walsh, W.M., 2013a. Evaluation of transposition and decomposition models for converting global solar irradiance from tilted surface to horizontal in tropical regions. Solar Energy 97, 369–387. http://dx.doi.org/10.1016/j.solener.2013.08.033. URL <http://www.sciencedirect.com/science/article/pii/S0038092X13003435> .
Yang, D., Dong, Z., Reindl, T., Jirutitijaroen, P., Walsh, W.M., 2014a. Solar irradiance forecasting using spatio-temporal empirical kriging and vector autoregressive models with parameter shrinkage. Solar Energy 103, 550–562. http://dx.doi.org/10.1016/j.solener.2014.01.024. URL <http://www.sciencedirect.com/science/article/pii/S0038092X14000425> .
Yang, D., Goh, G.S.W., Jiang, S., Zhang, A.N., 2016. Forecast UPC-level FMCG demand, Part III: Grouped reconciliation. In: 2016 IEEE International Conference on Big Data (Big Data), pp. 3813–3819, doi:http://dx.doi.org/10.1109/BigData.2016.7841053.
Yang, D., Goh, G.S.W., Jiang, S., Zhang, A.N., Akcan, O., 2015a. Forecast UPC-level FMCG demand, Part II: Hierarchical reconciliation. In: 2015 IEEE International Conference on Big Data (Big Data), pp. 2113–2121, doi:http://dx.doi.org/10.1109/BigData.2015.7363994.
Yang, D., Goh, G.S.W., Xu, C., Zhang, A.N., Akcan, O., 2015b. Forecast UPC-level FMCG demand, Part I: Exploratory analysis and visualization. In: 2015 IEEE International Conference on Big Data (Big Data), pp. 2106–2112, doi:http://dx.doi.org/10.1109/BigData.2015.7363993.
Yang, D., Gu, C., Dong, Z., Jirutitijaroen, P., Chen, N., Walsh, W.M., 2013b. Solar irradiance forecasting using spatial-temporal covariance structures and time-forward

kriging. Renew. Energy 60, 235–245. http://dx.doi.org/10.1016/j.renene.2013.05.030. URL <http://www.sciencedirect.com/science/article/pii/S0960148113002759> .

Yang, D., Jirutitijaroen, P., Walsh, W.M., 2012a. The estimation of clear sky global horizontal irradiance at the equator. Energy Proc. 25, 141–148. http://dx.doi.org/10.1016/j.egypro.2012.07.019. URL, PV Asia Pacific Conference 2011 <http://www.sciencedirect.com/science/article/pii/S1876610212011812> .

Yang, D., Jirutitijaroen, P., Walsh, W.M., 2012b. Hourly solar irradiance time series forecasting using cloud cover index. Solar Energy 86, 3531–3543. http://dx.doi.org/10.1016/j.solener.2012.07.029. URL, solar Resources <http://www.sciencedirect.com/science/article/pii/S0038092X12003039> .

Yang, D., Quan, H., Disfani, V.R., Liu, L., 2017b. Reconciling solar forecasts: Geographical hierarchy. Solar Energy 146, 276–286. http://dx.doi.org/10.1016/j.solener.2017.02.010. URL <http://www.sciencedirect.com/science/article/pii/S0038092X17301020> .

Yang, D., Quan, H., Disfani, V.R., Rodríguez-Gallegos, C.D., 2017c. Reconciling solar forecasts: Temporal hierarchy. Solar Energy 158, 332–346. http://dx.doi.org/10.1016/j.solener.2017.09.055. URL <http://www.sciencedirect.com/science/article/pii/S0038092X17308423> .

Yang, D., Sharma, V., Ye, Z., Lim, L.I., Zhao, L., Aryaputera, A.W., 2015c. Forecasting of global horizontal irradiance by exponential smoothing, using decompositions. Energy 81, 111–119. http://dx.doi.org/10.1016/j.energy.2014.11.082. URL <http://www.sciencedirect.com/science/article/pii/S0360544214013528> .

Yang, D., Walsh, W.M., Jirutitijaroen, P., 2014b. Estimation and applications of clear sky global horizontal irradiance at the equator. J. Solar Energy Eng. 136, 034505. http://dx.doi.org/10.1115/1.4027263.

Yang, D., Walsh, W.M., Zibo, D., Jirutitijaroen, P., Reindl, T.G., 2013c. Block matching algorithms: Their applications and limitations in solar irradiance forecasting. Energy Proc. 33, 335–342. http://dx.doi.org/10.1016/j.egypro.2013.05.074. URL, pV Asia Pacific Conference 2012 <http://www.sciencedirect.com/science/article/pii/S1876610213000842> .

Yang, D., Ye, Z., Lim, L.H.I., Dong, Z., 2015d. Very short term irradiance forecasting using the lasso. Solar Energy 114, 314–326. http://dx.doi.org/10.1016/j.solener.2015.01.016. URL <http://www.sciencedirect.com/science/article/pii/S0038092X15000304> .

Yang, D., Ye, Z., Nobre, A.M., Du, H., Walsh, W.M., Lim, L.I., Reindl, T., 2014c. Bidirectional irradiance transposition based on the Perez model. Solar Energy 110, 768–780. http://dx.doi.org/10.1016/j.solener.2014.10.006. URL <http://www.sciencedirect.com/science/article/pii/S0038092X14004927> .

Yang, H., Kleissl, J., 2016. Preprocessing WRF initial conditions for coastal stratocumulus forecasting. Solar Energy 133, 180–193. http://dx.doi.org/10.1016/j.solener.2016.04.003. URL <http://www.sciencedirect.com/science/article/pii/S0038092X16300251> .

Yang, H., Kurtz, B., Nguyen, D., Urquhart, B., Chow, C.W., Ghonima, M., Kleissl, J., 2014d. Solar irradiance forecasting using a ground-based sky imager developed at UC San Diego. Solar Energy 103, 502–524. http://dx.doi.org/10.1016/j.solener.2014.02.044. URL <http://www.sciencedirect.com/science/article/pii/S0038092X14001327> .

Yang, H.T., Huang, C.M., Huang, Y.C., Pai, Y.S., 2014e. A weather-based hybrid method for 1-day ahead hourly forecasting of PV power output. IEEE Trans. Sustain. Energy 5, 917–926. http://dx.doi.org/10.1109/TSTE.2014.2313600.

Yang, Z., 2015. Large-eddy simulation: Past, present and the future. Chin. J. Aeronaut. 28, 11–24. http://dx.doi.org/10.1016/j.cja.2014.12.007. URL <http://www.sciencedirect.com/science/article/pii/S1000936114002064> .

Ye, J., 1998. On measuring and correcting the effects of data mining and model selection. J. Am. Statist. Assoc. 93, 120–131. http://dx.doi.org/10.1080/01621459.1998.10474094.

Zagouras, A., Pedro, H.T.C., Coimbra, C.F.M., 2015. On the role of lagged exogenous variables and spatio-temporal correlations in improving the accuracy of solar forecasting methods. Renew. Energy 78, 203–218. http://dx.doi.org/10.1016/j.renene.2014.12.071. URL <http://www.sciencedirect.com/science/article/pii/S0960148115000051> .

Zamo, M., Mestre, O., Arbogast, P., Pannekoucke, O., 2014a. A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production, Part I: Deterministic forecast of hourly production. Solar Energy 105, 792–803. http://dx.doi.org/10.1016/j.solener.2013.12.006. URL <http://www.sciencedirect.com/science/article/pii/S0038092X13005239> .

Zamo, M., Mestre, O., Arbogast, P., Pannekoucke, O., 2014b. A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production. Part II: Probabilistic forecast of daily production. Solar Energy 105, 804–816. http://dx.doi.org/10.1016/j.solener.2014.03.026. URL <http://www.sciencedirect.com/science/article/pii/S0038092X14001601> .

Zhang, J., Florita, A., Hodge, B.M., Lu, S., Hamann, H.F., Banunarayanan, V., Brockway, A.M., 2015a. A suite of metrics for assessing the performance of solar power forecasting. Solar Energy 111, 157–175. http://dx.doi.org/10.1016/j.solener.2014.10.016. URL <http://www.sciencedirect.com/science/article/pii/S0038092X14005027> .

Zhang, J., Hodge, B.M., Lu, S., Hamann, H.F., Lehman, B., Simmons, J., Campos, E., Banunarayanan, V., Black, J., Tedesco, J., 2015b. Baseline and target values for regional and point PV power forecasts: Toward improved solar forecasting. Solar Energy 122, 804–819. http://dx.doi.org/10.1016/j.solener.2015.09.047. URL <http://www.sciencedirect.com/science/article/pii/S0038092X1500540X> .

Zhang, Q., Deng, X., Qian, P.Z.G., Wang, X., 2013. Spatial modeling for refining and predicting surface potential mapping with enhanced resolution. Nanoscale 5, 921–926. http://dx.doi.org/10.1039/C2NR33603K.

Zhang, X., 2011. Structural risk minimization. In: Sammut, C., Webb, G.I. (Eds.), Encyclopedia of Machine Learning. Springer, New York, pp. 929–930.

Zhang, Y., Li, X., Bai, Y., 2015c. An integrated approach to estimate shortwave solar radiation on clear-sky days in rugged terrain using MODIS atmospheric products. Solar Energy 113, 347–357. http://dx.doi.org/10.1016/j.solener.2014.12.028. URL <http://www.sciencedirect.com/science/article/pii/S0038092X14006197> .

Zhong, X., Kleissl, J., 2015. Clear sky irradiances using REST2 and MODIS. Solar Energy 116, 144–164. http://dx.doi.org/10.1016/j.solener.2015.03.046. URL <http://www.sciencedirect.com/science/article/pii/S0038092X15001735> .

Zhu, T., Wei, H., Zhao, X., Zhang, C., Zhang, K., 2017. Clear-sky model for wavelet forecast of direct normal irradiance. Renew. Energy 104, 1–8. http://dx.doi.org/10.1016/j.renene.2016.11.058. URL <http://www.sciencedirect.com/science/article/pii/S0960148116310424> .