Solar Energy 162 (2018) 232–247

Contents lists available at ScienceDirect

Solar Energy

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

A comparative study of LSTM neural networks in forecasting day-ahead global horizontal irradiance with satellite data

Shikhar Srivastava*, Stefan Lessmann

School of Business and Economics, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany

ARTICLE INFO

Keywords:
Solar energy forecasting
Long short term memory
Deep learning
Remote sensing data

ABSTRACT

Accurate forecasts of solar energy are important for photovoltaic (PV) based energy plants to facilitate an early participation in energy auction markets and efficient resource planning. The study concentrates on Long Short Term Memory (LSTM), a novel forecasting method from the family of deep neural networks, and compares its forecasting accuracy to alternative methods with a proven track record in solar energy forecasting. To provide a comprehensive and reliable assessment of LSTM, the study employs remote-sensing data for testing predictive accuracy at 21 locations, 16 of which are in mainland Europe and 5 in the US. To that end, a novel framework to conduct empirical forecasting comparisons is introduced, which includes the generation of virtual PV plants. The framework enables richer comparisons with higher coverage of geographical regions. Empirical results suggest that LSTM outperforms a large number of alternative methods with substantial margin and an average forecast skill of 52.2% over the persistence model. An implication for energy management practice is that LSTM is a promising technique, which deserves a place in forecasters' toolbox. From an academic point of view, LSTM and the proposed framework for experimental design provide a valuable environment for future studies that assess new forecasting technology.

1. Introduction

By 2035, the demand for primary energy and electricity is projected to increase by 50% and 70%, respectively (Duffy et al., 2015). In the face of rising threats of fossil fuel based dependencies, the International Energy Agency (IEA) encourages research on renewable energy related technologies for a sustainable future. Solar and wind energy are the two most popular renewable energy sources. Especially solar energy is an abundant but underutilized source of energy (Perez et al., 2011). Considering technological progress leading to higher PV efficiency, economies of scale, and the further development of energy markets, a substantially higher contribution of solar energy to future energy mix can be expected (Sangster, 2014).

In many countries, energy markets are divided in day-ahead and intra-day market sessions. For power plants, it is necessary to supply a required amount of energy. Based on this requirement and predictions whether the quote is likely to be met, plant authorities make decisions regarding energy purchases. The sooner a commitment is made, the lower the price. Considering that markets follow a bid system, this implies that larger forecast errors will lead to higher cost of energy. The difficulty in efficient market penetration for solar energy lies in its high variability and uncertainty in both time and space. Therefore, accurate forecasting of energy availability remains a top priority in the industry. Due to data policies of local solar energy companies, PV data is often unavailable in which case scientists tend to perform energy predictions via forecasting Global Horizontal Irradiance (GHI).

Standard practice in solar energy forecasting includes use of mesoscale numerical weather prediction (NWP) models, which are built upon global forecasts of atmospheric variables. Such models have evolved over decades. Although their ability to explain intrinsic relationships has earned them industry-wide trust, their complexity hinders further improvement in forecasting (Larson, 2013). On the other hand, models using supervised machine learning (ML) are becoming increasingly popular and often show promising accuracy (Kleissl, 2013). Even some of the NWP mesoscale models use ML methods to improve their predictions. For example, BLUE Forecast, a mesoscale NWP model built upon Global Forecast System (GFS) outputs, integrates a statistical post-processing procedure, which uses different learning methods such as ridge regression, or neural networks (Perez et al., 2013). ECMWF-Ol, another popular NWP model, also uses neural networks for post-processing NWP outputs (Lorenz et al., 2009b). When coupled with on-going evolution of computational abilities and increased accuracy of GFS forecasts, this flexibility has created a surge of research over the last decade.

* Corresponding author.
E-mail addresses: shikhar.srivastava@outlook.com (S. Srivastava), stefan.lessmann@hu-berlin.de (S. Lessmann).

https://doi.org/10.1016/j.solener.2018.01.005
Received 11 August 2017; Received in revised form 4 December 2017; Accepted 3 January 2018

Despite a large body of literature on GHI forecasting, experiences with deep learning techniques, the latest trend in the machine learning, are yet scarce (Inage, 2017). Moreover, due to the geo-temporal nature of solar energy modeling, we observe a lack of frameworks to robustly compare alternative forecasting methods. The paper contributes to literature through exploring a relatively new learning method, Long Short Term Memory (LSTM) for day-ahead solar energy forecasting. LSTM belongs to the family of deep neural networks, which have shown excellent performance in a variety of applications such as computer vision, text analysis, and many others (LeCun et al., 2015). However, to the best of our knowledge, this is the first study to test the potential of LSTM for solar energy forecasting. Capitalizing on the availability of high quality remote-sensing data, a second contribution of the paper consists of a novel experimental framework that facilitates validation of forecasting models in varying climatic/topographical conditions through introducing virtual solar stations. The framework increases the breadth of empirical comparisons and allows comparing alternative methods under different conditions, which helps to secure reliability of conclusions related to the relative suitability of alternative forecasting methods (Armstrong and Fildes, 2006).

The paper is organized as follows: Section 2 reviews previous literature. Section 3 elaborates on methodology and introduces LSTM. Sections 4–7 explain the data sources and experimental setup together with its underlying motivations. A discussion of empirical results is available in Section 8, before the paper concludes in Section 9.

2. Related work

Irradiance forecasting does not follow a one-fits-all approach. Especially the forecasting horizon determines the suitability of alternative models (Kostylev and Pavlovski, 2011). For example, to support decision-making in operational management (load in-/decrement, peak load matching, etc.), prior research has studied short-term models that forecast irradiance from 30 min to a few hours ahead (Chow et al., 2011; Mathiesen and Kleissl, 2011). The focus of this paper is one day ahead forecasting. A daily horizon is established in the literature and meaningful from an economic perspective because increases in the accuracy of day-ahead solar energy forecasts may facilitate substantial cost savings (Brancucci Martinez-Anido et al., 2016). Candidate applications for day-ahead forecasts include load dispatch planning, making unit commitments and maintaining operational security (Voyant et al., 2017; Soman et al., 2010; Wang et al., 2011).

In the following, we discuss prior work with comparable forecast horizon. More specifically, we review studies that predict GHI using ML algorithms with forecast horizon of one day or above. Diagne et al. (2013), Inman et al. (2013) and Raza et al. (2016) provide comprehensive reviews of other solar energy forecasting methods.

Designing a GHI forecasting benchmark offers several degrees of freedom and these complicate cross-study comparisons. For example, many studies employ data with specific spatio-temporal context. Given the unique characteristics of weather in both space and time, there is no guarantee that a method, which has been found successful at one location is as effective at a different location. To depict the state-of-the-art in daily solar energy forecasting, Table 1 classifies prior studies along dimensions, the corresponding data, and the employed forecasting techniques.

Table 1 reveals that few previous studies involve an evaluation of models across several locations or different countries. Exceptions with larger scope include Linares-Rodríguez et al. (2011), who, after getting access to radiation data of 83 stations of a small coastal region in Spain, develop a regional forecasting model. Moreno et al. (2011) devise a neural network model for the whole country of Spain using ground readings recorded at 40 stations sparsely distributed across the country.

With respect to forecasting methods, Table 1 indicates that feed forward neural networks (FFNN) prevail in irradiation forecasting. Some authors have also considered more advanced network architectures. For example, to capture non-linearity in time-series of weather related variables, Cao and Lin (2008) propose diagonal recurrent wavelet neural networks for solar irradiance forecasting.

In contrast to neural network based models, a supervised ML algorithm called gradient boosting regression (GBR), which has received little attention in the domain, has won a highly visible solar-energy prediction competition hosted on Kaggle[1] in 2014 (McGovern et al., 2015). The competition was organized by the American Meteorological Society (AMS) and provided a huge dataset containing 14 years of temporal records and PV output values from 98 PV plants covering the whole region of Oklahoma. The data includes forecasts of atmospheric variables at different time steps of a day from a global model. Solar energy data of such magnitude in both time and space has rarely been examined before.

Arguably, the result of the AMS competition has further increased interest in evaluating modern ML techniques in the mid-range irradiation forecasting literature. Several such techniques have been tested in recent work. For example, Junior et al. (2014) and Ekici (2014) obtain promising results using support vector regression (SVR) for irradiance forecasting. Li et al. (2016) and Gala et al. (2016) introduce ensemble models. The former combine FFNN and GBR, whereas the latter demonstrate how an ensemble of SVR, GBR and random forest regression forecast irradiance more accurately than any of these (base) models in isolation.

Multiple conclusions emerge from Table 1. First, the potential of deep learning and LSTM in particular for irradiation forecasting has not been examined. Voyant et al. (2017), in a recent review on irradiance forecasting using ML methods, arrive at a similar conclusion. Second, no study considers more than one country. This might be due to prior studies depending on ground readings of irradiance. When concentrating on a single country, it may be difficult to accommodate inter-climatic validations of models. Therefore, a restriction of many previous results is that they only explain the irradiation behavior of a specific location.

In this paper, we strive to overcome these issues. We address bias toward location and climate specific modeling through using satellite based GHI measurements. Satellite based GHI measurements are recorded for many geo-locations. The availability of corresponding data facilitates a broader comparison of forecasting models across multiple locations and climate conditions.

In particular, we address bias toward location and climate specific modeling through using satellite based GHI measurements. This allows us to examine the accuracy of forecasting models at multiple locations. In addition, we use a full year of data for model validation to test model effectiveness across all seasons.

A second objective of the study is to introduce LSTM to the community and tests its potential for irradiance forecasting. As shown in Table 1, examining new forecasting technology is an established research goal (Voyant et al., 2017). We acknowledge that studies which pursue this goal face a risk to over-tune the focal forecasting method (Hand, 2006). Avoiding reliance on ground measurements and facilitating multi-location analysis, our evaluation framework allows mitigating this risk in that we avoid excessive tuning of forecasting models to individual locations but re-use model configuration for multiple locations. With respect to benchmark methods, we select other ML algorithms with proven track record in irradiance forecasting, namely GBR and FFNN. We also consider Persistence as a standard benchmark.

3. Long Short Term Memory (LSTM)

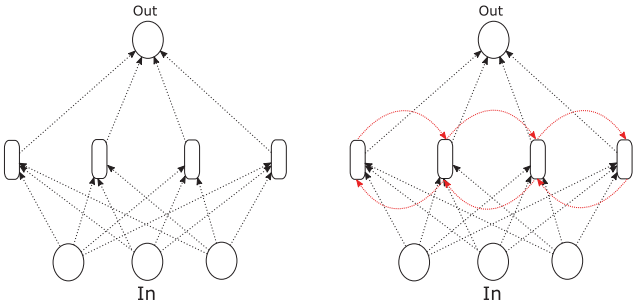
The LSTM algorithm belongs to the family of (deep) recurrent neural networks (RNN). As shown in Fig. 1, RNN differ from FFNN, one

[1] Kaggle is an online platform for data mining competitions. Competition website: http://www.kaggle.com/c/ams-2014-solar-energy-prediction-contest.

Table 1
Previous work on irradiance forecasting with horizon of at least 24 h.

Study (in chronological order)	Geographical span			Data		Method[c]	
	Countries covered	Distinct locations	Climates covered[a]	Training set size[b]	Test set size[b]	Proposed	Benchmark
Kemmoku et al. (1999)	1	1	1	1827	365	FFNN	Persistence
Podestá et al. (2004)	1	1	1	365	1276	GAM	None
Yona et al. (2007)	1	1	1	333	4	RNN, RBFNN	FFNN
Cao and Lin (2008)	1	1	1	7992	744	DRWNN	AP, FFNN
Bacher et al. (2009)	1	1	1	1460	NA	RLS	Persistence
Paoli et al. (2010)	1	1	1	6574	547	FFNN	NP, MC, K-NN, AR
Linares-Rodríguez et al. (2011)	1	83	1	219,439	60,768	FFNN	None
Marquez and Coimbra (2011)	1	1	1	236	158	FFNN	NWP, Persistence
Ding et al. (2011)	1	1	1	7	2	FFNN	None
Chen et al. (2011)	1	1	1	670	74	RBFN	None
Moreno et al. (2011)	1	40	3	1241	219	FFNN	BC, KRR
Cococcioni et al. (2011)	1	1	1	359	6	FFNN	None
Ekici (2014)	1	1	1	1096	365	SVM	None
Junior et al. (2014)	1	4	2	1464	1460	SVM	Persistence
Voyant et al. (2014)	1	1	1	2922	731	FFNN	ARMA, Persistence
Salcedo-Sanz et al. (2014)	1	1	1	292	73	Hybrid ELM	ELM, SVR
Long et al. (2014)	1	1	1	723	38	None	FFNN, KNN, MLR
Fernández et al. (2014)	1	1	1	365	181	SVR	None
Cornaro et al. (2015)	1	1	1	876	584	FFNN	Persistence
Li et al. (2016)	1	1	1	365	NA	FFNN	SVR
Gala et al. (2016)	1	7	2	365	302	Ensemble	SVR, GBR, RFR
Wang et al. (2016)	1	1	1	365	181	PF-LR	FFNN, RBF
Leva et al. (2017)	1	1	1	120	30	FFNN	None
This study[d]	7	17	10	164,350	365	LSTM	GBR, FFNN, Persistence

[a] Variation in climate follows Koeppen Climate Classification System (Peel et al., 2007).
[b] Training and test sizes are number of distinct days of taking the observations regardless of further granular measurements.
[c] Abbreviations have the following meaning: FFNN = Feed Forward Neural Network, GBR = Gradient Boosting Regression, LSTM = Long Short Term Memory, GAM = Generalized Additive Model, RBF = Radial Bias Function, DRWNN = Diagonal Recurrent Wavelet Network, RLS = Recursive Least Square, AP = ångström-Prescott equation, K-RR = Kernel Ridge Regression, NP = Naïve Predictor, MC = Markov Chain, K-NN = K-Nearest Neighbor, AR = Auto-Regression, BC = Bristow-Campbel, ARMA = Autoregressive Moving Average, ELM = Extreme Learning Machine, SVR = Support Vector Regression, RFR = Random Forest Regression, PF-LR = Partial Functions-Linear Regression.
[d] For each distinct location, we create a training region of 50 virtual stations. The train and test sizes are per location.



(a) A simple FFNN (b) A simple RNN

Fig. 1. Bifurcation of neural networks.

of the most popular forecasting methods in solar energy forecasting (Table 1), in that they incorporate feedback loops. FFNNs were conceptually developed in the early-1940s inspired by the interactive design of nervous system in human brain. In this respect, RNNs can be seen as an advancement over FFNNs and LSTMs over RNNs.

The architecture of RNN allows hidden units to share parameters across time indexes. In theory, this sharing effectively helps in building a "memory" of long sequences, allowing the system to recognize and predict sequences. Arguably, RNN have a closer relationship to the functioning of synapses in the human brain than FFNN in that humans typically learn in progressive rather than randomized order. Feedback loops enable a RNN to learn progressively. For example, a single neuron can be considered an information processing unit. The feedback loop provides the unit with memory. The previous state of the neuron (e.g., information from a previous time step) can be re-incorporated and taken into account when updating the memory.

In practice, however, the vanishing gradient problem prohibits application of this concept when learning from deep sequences or from large contexts. As has been thoroughly explained in Bengio et al. (1994), the problem arises from propagation of local errors at each time index. Even our brains are incapable of recalling every single event from the past.

LSTM addresses the vanishing gradient problem through incorporating self-connected "gates" in the hidden units. When working over longer periods of time, the gates allow LSTM units to read, write, and remove information from memory. This characteristic enables the units to hold only relevant data while 'forgetting' irrelevant information, and hence, sustain a constant error Hochreiter et al. (2001). This, again, mimics the way in which human selectively remember details from the past. For GHI forecasting, we expect LSTM to exploit the temporal and spatial dependence of data, which represents a major asset in utilizing contextual information. LSTM and its adaptations have shown potential to solve a range of problems involving sequential learning (Graves, 2012). In recent years, there has been a dramatic increase in LSTM development and use. Arguably, this increase can be attributed to the evolution of GPU based data processing units allowing efficient computation of large number of hidden units embedded in multi layered networks.

3.1. LSTM calculations

To understand the functioning of LSTM memory units, Fig. 2 depicts a single localized LSTM cell in the first layer of a network at a time step t. Input (i_t), update (g_t), output (o_t) and forget (f_t) represent the output values of the four types of gates. The gates receive an input of the same LSTM unit's output obtained at a previous time step (h_{t-1}). The gates also receive input data related to the current time step (x_t). In Fig. 2, we use colored arrows to depict this message flow. Eqs. (1)–(6) are

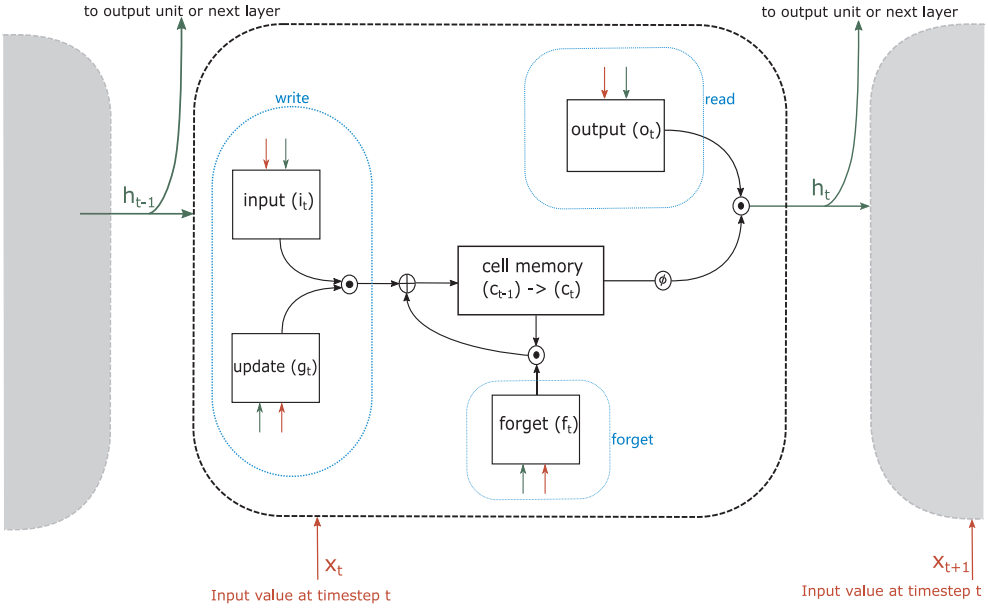


Fig. 2. An LSTM unit in the 1st layer for time step t. The symbols ⊙ and ⊕ represent point-wise scalar multiplication and the sum function, respectively. Color coded arrows show the direction of inputs to a system. The cell output activation function, represented by ∅, inside a O is user-defined and set to *ReLu* in this experiment. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

governing equations, which we explain below. Please note that θ_{xn} and θ_{hn} are weight matrices of input and output data from previous cells, whereas b_n represents bias vectors with $n \in (i,g,f,u)$, which identify the aforementioned gates. As is standard practice in neural network learning, the parameters are initialized through sampling from a uniform distribution and updated using back-propagation technique (Schmidhuber, 2015).

The forget gate, which is a key element of the LSTM architecture, chooses how much information to delete from memory. It operates through a *hard sigmoid* activation (linearly approximated standard *sigmoid*) which has a range of 0 (forget all) to 1 (remember all) and is governed by (1).

$$f_t = sigm(\theta_{xf} x_t + \theta_{hf} h_{t-1} + b_f) \tag{1}$$

The interaction (element-wise multiplication) of input and update gate (see Fig. 2) performs write function into memory. The input gate, using *hard sigmoid*, decided which values to write and the update gate, using *hyperbolic tangent* (*tanh*) activation, creates a vector of new cell values. Eqs. (2) and (3) represent their working. A near zero value of input activation will stop new inputs from updating cell memory and hence, stored values can be used till much later.

$$i_t = sigm(\theta_{xi} x_t + \theta_{hi} h_{t-1} + b_i) \tag{2}$$

$$g_t = tanh(\theta_{xg} x_t + \theta_{hg} h_{t-1} + b_g) \tag{3}$$

Cell memory (see Fig. 2) updates itself recursively by interaction of its old value ($t-1$) with forget and write gates' values. It follows (4)

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \tag{4}$$

Finally, the output gate, which implements a read function, is combined (element-wise multiplication) with cell memory to calculate the output of the cell h_t, also referred as the cell state. An activation function, ϕ, is performed on the cell memory to bind output values, and to determine which values will go in the output from the output gate, o_t. The activation function, ϕ, in the cell output equation is often *tanh*. However, for this experiment, we use *Rectified Linear Unit* (*ReLu*), which is a more popular choice in recent deep learning literature (LeCun et al., 2015). Eq. (5) governs the read function of the output gate o_t and (6) calculates the final cell output.

$$o_t = sigm(\theta_{xo} x_t + \theta_{ho} h_{t-1} + b_o) \tag{5}$$

$$h_t = o_t \odot \phi(c_t) \tag{6}$$

3.1.1. Regularization

LSTMs or deep neural networks in general are prone to over-fitting. As the network learns, weights adjust so much that they not only explain structural patterns in the data but also noise and idiosyncrasies specific to the training data. We protect against over-fitting using dropout regularization, which has become common practice in deep learning (Szegedy et al. (2015)). A dropout layer blocks a random set of cell units in one iteration of LSTM training. Blocked units do not receive and do not transmit information. Removing connections in the network reduces the number of free parameters to be estimated during training and the complexity of the network. Consequently, dropout helps to prevent over-fitting.

The fraction of blocked network cells (dropout layer value) is a user-defined parameter and commonly set using default values or cross-validation (Srivastava et al., 2014; Szegedy et al., 2015) Networks that include a dropout layer can be considered an ensemble of multiple networks with each ensemble member being created in one training iteration via bagging of the unblocked units. Fig. 3 shows an instance (one iteration scenario), with a dropout layer applied after two fully connected layers (rectangular figures). In one iteration, a randomly selected fraction of 20 percent of the total number of units are blocked and do not participate in network training. Thus, a different network is trained in each iteration. When calculating the final network prediction, one implicitly averages over the different networks devised during training, which is equivalent to constructing an ensemble forecast. Much literature provides theoretical and empirical results that explain and confirm the success of ensemble modeling to increase forecast accuracy (Caruana and Niculescu-Mizil, 2006; Nanculef et al., 2006; Banfield et al., 2005). The ensembling mechanism of is another reason

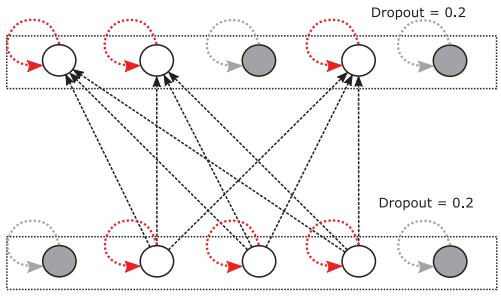


Fig. 3. A dropout layer applied after two fully connected hidden layers.

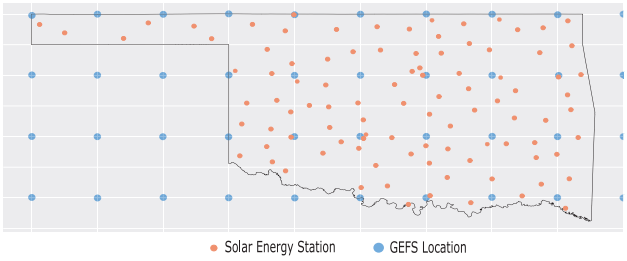


Fig. 4. Geo-Spread of stations across Oklahoma in the AMS competition data.

why dropout often increases predictive accuracy.

4. Data engineering

The 2014 AMS Kaggle competition on solar energy forecasting provides daily energy production data for a total of 98 stations, which are spread throughout the state of Oklahoma in an irregular fashion. The competition task is to forecast day ahead energy based on day ahead prediction of atmospheric variables using NWP based 2nd generation NOAAs Global Ensemble Forecast System (GEFS) Re-forecast weather data. Fig. 4 shows the stations' locations on a map of Oklahoma. The competition data is available for public access and researchers are still allowed to submit forecasts for comparison purposes. The large variance in topography and climate made the AMS competition a valuable environment to test the efficiency of forecasting methods. In this study, we employ the AMS for an initial testing of LSTM and to identify a suitable LSTM configuration. To that end, we design an experiment that mimics the original AMS competition with some deviations (see below). Importantly, the task and structure of the AMS competition also inspired the approach to construct virtual stations (see below), which allows for a broad and robust evaluation of forecasting methods at different geographic regions. The following sections revisit the AMS competition, introduce the experimental data, and elaborate on the virtual station construction approach.

4.1. Input

Table 2 depicts the input variables in the AMS data set. In this study, we further augment the data with three variables (first three rows of Table 2) that can be created from the information available in the competition data. Details of these inputs are provided below.

4.1.1. GEFS data

The NOAA's Reforcast weather data of the AMS competition allows for a resolution of 1° × 1° latitude and longitude of the planet. The model predicts various atmospheric variables (including solar irradiation) using physical modeling techniques (Lorenz and Heinemann, 2012). The prediction is made for a time step of three hours for the next three days. Thereafter, another five days are forecast using a time step of six hours. For each atmospheric variable, 11 ensemble members - initialized with small perturbations - are created.[2] Considering that the interaction of atmospheric variables along with its own trend could be highly relevant for day ahead irradiation prediction, we consider seven time steps i.e. from 6 to 24 h (one every three hours). Note that the AMS competition considered five time steps (from 12 to 24 h).

Table 2
List of input variables.

Variable	Unit
Haversine distance between GEFS (x) and station location	km
Difference between elevation of GEFS (x) and station location	km
Day of the year	
3-Hour accumulated precipitation at the surface	$kg\,m^{-2}$
Downward long-wave radiative flux average at the surface	$W\,m^{-2}$
Downward short-wave radiative flux average at the surface	$W\,m^{-2}$
Air pressure at mean sea level	Pa
Precipitable Water over the entire depth of the atmosphere	$kg\,m^{-2}$
Specific Humidity at 2 m above ground	$kg\,kg^{-1}$
Total cloud cover over the entire depth of the atmosphere	%
Total column-integrated condensate over the entire atmos.	$kg\,m^{-2}$
Maximum Temperature over the past 3 h at 2 m above the ground	K
Minimum Temperature over the past 3 h at 2 m above the ground	K
Current temperature at 2 m above the ground	K
Upward long-wave radiation at the surface	$W\,m^{-2}$
Upward long-wave radiation at the top of the atmosphere	$W\,m^{-2}$
Upward short-wave radiation at the surface	$W\,m^{-2}$

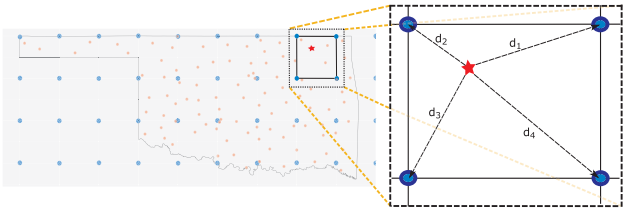


Fig. 5. A schematic view of single grid surrounding the location.

The blue dots in Fig. 4 are points for which GEFS data is available. The task is to create a mesoscale model using information of these dots near the solar stations and minimize the overall prediction error. Aler et al. (2015), who also experiment with the AMS data, show that there is little improvement in forecast accuracy when considering more than four GEFS points surrounding a solar energy station. Also, the summary of the results of the AMS competition suggests considering four corners of a square shaped grid surrounding the solar station to be a suitable choice (McGovern et al., 2015). Fig. 5 illustrates this approach. The square is shaped of 1° × 1° latitude and longitude. We follow this approach in our experiment and employ data of the four GEFS points cornering a solar station (location in focus). Recall that most input variables (Table 2) are GEFS sourced atmospheric data. We average over the 11 ensemble member per atmospheric variable to obtain a single value for one time step. Then, the final inputs for each location under study includes seven values (one per time step from 6 to 24 h) per atmospheric variable.

4.1.2. Spatial/Temporal data

The atmospheric variable values captured at the corners surrounding a station aid a forecast model to learn spatial structure. In addition, we propose the first three entries of Table 2 as auxiliary geo/temporal input to a forecasting model. First, we calculate the distance between a focal location and its GEFS corner points using the Haversine Distance formula (Robusto, 1957) (see Fig. 5). A second auxiliary input is the difference between the elevations of the GEFS corner points and a focal location. We obtain the elevation of the focal location from the Google Maps API.[3] Finally, we consider the day of the year as model input.

4.2. Target variable

Solar energy data requires a PV setup. Unfortunately, industry rarely provides corresponding data for public access. Data sets similar to that of AMS competition are thus difficult to obtain. An alternative method is to predict GHI instead. Lorenz et al. (2010) explain a GHI to PV conversion procedure.The common way of recording GHI is via installation of Pyranometers on a planar surface. However, in terms of geographical span, Pyranometer readings have limited availability due to installation and constant maintenance costs. Furthermore, other

[2] The dataset is available for public access, instructions can be found in Hamill et al. (2013).
[3] We use the Python-based "geocoder 1.23.2" module to gather elevation data.

quality problems may cause unexpected large GHI readings at night and small readings in day duration. Correction of these readings to real values should ideally be done by the local station manager, which often is not possible when working with retro-data.

Given these issues with GHI readings, several scholars suggest approximating GHI readings using satellite images (Lorenz et al., 2009b; Lorenz and Heinemann, 2012; Tadesse et al., 2016; Sengupta et al., 2015). USA based National Solar Radiation Database (NSRDB) allows free access to Physical Solar Model (PSM) and Europe's Copernicus Atmosphere Monitoring Service (CAMS) to its Radiation Service.[4] This dataset allows a spatial resolution of 4 × 4 km with a temporal resolution of up to 30 min. It is important to note that NSRDB and CAMS data are not forecasts. To verify feasibility of approximating ground readings using satellite based data, we compute the correlation between the original target variable in the AMS competition, which is cumulative daily PV output of the 98 stations (Fig. 4), and a satellite image based proxy for every station. This produces Pearson correlation coefficients of 0.95–0.99 for every year during 1998–2007. Analog correlation values are claimed to be between 0.91 and 0.97 for the CAMS Radiation service (Schroedter-Homscheidt, 2016).

It is noteworthy that satellite based GHI values suffer higher measurement error compared to pyranometer readings (Nottrott and Kleissl, 2010). This constraints the use of satellite-derived data. In particular, forecast models based on satellite based GHI values cannot be validated to a level of accuracy higher than that of the measurements. However, the aim of this study is to provide a relative estimate of the accuracy of alternative forecasting models and LSTM in particular. Using satellite based GHI values as target variable is valid in this setting because the methods under study predict the same target variable from the same input variables. Methods are thus compared on even ground. The major advantage of satellite data is its ease of access for various geo-locations. This facilitates a multi-location validation of forecasting models and promises more robust results compared to testing models in a single location. In addition, there are no missing values in the data. When the purpose of forecasts is to inform bidding decisions in solar-energy bidding market, satellite-derived data can be combined with data from ground-readings; using the latter whenever more accurate ground readings are available and satellite based proxies otherwise.

4.3. Region construction

Initial testing with the AMS competition data reveals that a regional modeling approach performs better than developing forecasting models for specific locations. This agrees with traditional methods of mesoscale NWPs that generally take weighted/pixel averages of predicted variables based on the distance of a focal location from GEFS points. Taking advantage of vast and detailed spatial availability of the satellite based proxy target variable, we design a learnable region within the small 1° × 1° grid. Inspired from the nature of the AMS competition data and initial results concerning regional forecasting outperforming location specific modeling, we sample a set of what we call virtual solar energy stations inside the grid. The underlying idea is to create a grid level model where the algorithm learns the spatial and temporal behavior of the region pertaining to irradiance received in it. Fig. 6 illustrates the approach for a hypothetical region with one station (red star) and the corresponding GEFS corner points. We create a total of 49 virtual stations, 36 of which are uniformly distributed (at 0.2° from horizontally or vertically placed neighboring virtual stations). The remaining 13 are created randomly. With the given resolution of satellite data, we argue that it is safe to build the uniformly distributed virtual stations at such intervals whilst avoiding introducing bias. Including the station to

[4] For NSRDB's PSM, readers can refer to Sengupta et al. (2012) for methodological details and validation of calculated against real GHI values.

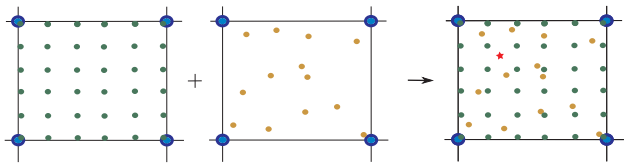


Fig. 6. Virtual station construction - 36 uniformly and 13 randomly distributed. Star represents station in focus.

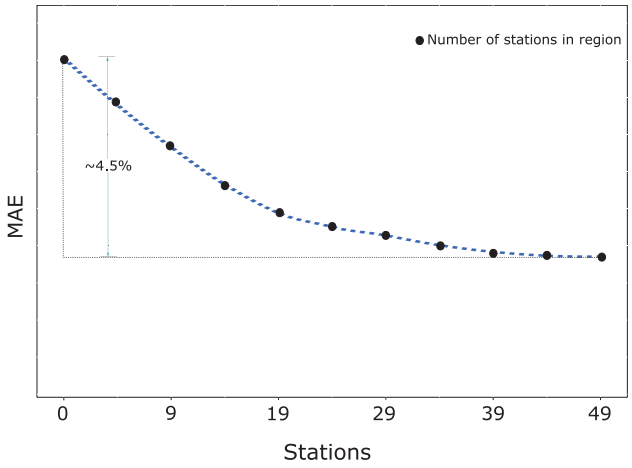


Fig. 7. Average reduction of forecast error in MAE when increasing the number of stations in a region.

forecast, a learning algorithm can draw on data from 50 stations to learn a forecasting model. When testing with real stations of the AMS data, we observe a gradual increase in forecast accuracy when increasing the number of virtual stations. Fig. 7 provides corresponding results in terms of MAE. The figure motivates the choice of 50 stations in that a further increase of the number of stations does not offer an improvement (i.e., does not reduce MSE further). An important advantage of the virtual station approach is its replicability and ease of implementation. This input architecture allows evaluating forecasting model performance over any part of the world where high resolution satellite based GHI values are recorded.

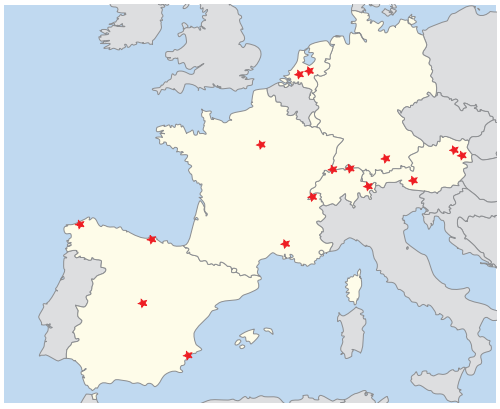
Lorenzo et al. (2015) showed that spatial averaging is a potential source of underestimation of forecast variability. However, results from Lorenzo et al. (2015) also suggest a worsening of the trade-off between skill and variance with increase in forecast horizon. The study was performed for models with horizon of 1–120 min. The variance of their proposed model converged to that of a spatially-averaged model for more than 30 min forecast horizon. We consider a 24 h horizon in our study. This substantial difference suggests that the increase in variance from regional modeling might be negligible in our setting. Empirical evidence derived from initial testing on AMS data supports this view. Also, we expect spatial averaging to affect alternative forecasting models in a similar manner, so that the comparison of LSTM to benchmark models is largely unaffected from spatial averaging.

4.4. Validation

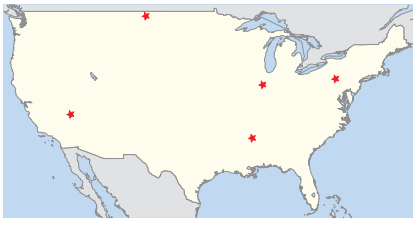
From the nature of solar irradiation forecasting using global-scale models, dependencies on location specific climatic conditions and geospatial spread are conspicuous. We suggest that an objective validation and comparison of forecasting methods should include multi-location based experiments. A diverse set of climatic conditions and topographical properties across different locations increases confidence in observed results and improves the external validity of empirical studies. With this in mind, we test LSTM in a forecasting benchmark that encompasses 5 locations from the United States and 16 locations from six

Table 3
Locations and their climate types.

Country/Location	Latitude	Longitude	Elevation (m)	Climate
Germany				
Hohenpeissenberg	47.8	11.01	925	Continental
Netherlands				
Cabauw	51.97	4.93	−1	Humid continental
De Bilt	52.1	5.18	2	Humid continental
Austria				
Grossenzersdorf	48.2	16.57	154	Continental
Sonnblick	47.05	12.96	3013	Alpine
Wien	48.25	16.35	215	Continental
France				
Carpentras	44.08	5.06	100	Mediterranean
Palaiseau	48.71	2.21	158	Continental
Spain				
Santander	43.48	−3.8	69	Oceanic
La Coruna	43.35	−8.42	43	Oceanic
Alicante	38.28	−0.55	18	Mediterranean/ Humid
Madrid	40.45	−3.73	634	Continental
Switzerland				
Basel	47.55	7.58	273	Temperate Atlantic
Zurich	47.48	8.53	427	Moderate maritime/ continental
Geneva	46.25	6.13	419	Moderate maritime/ continental
Davos	46.82	9.85	1559	Continental/Alpine
USA				
Desert Rock	36.63	−116.02	1007	Arid
Bondville	40.07	−88.37	213	Continental
Fort Peck	48.32	−105.10	634	Continental
Goodwin Creek	34.25	−89.87	98	Humid continental
Penn State	40.72	−77.93	374	Humid continental



(a) Europe based locations



(b) USA based locations

Fig. 8. Locations on map.

different countries of mainland Europe. Table 3 provides information on the selected locations. As can be seen from Figs. 8(a) and (b), we chose locations so as to achieve a high diversity. For each location, we create virtual stations to obtain input data for forecast model development (see Section 4.3). Although ground recorded GHI data is publicly available for some of the US locations considered in the study (i.e., SURFRAD locations), we refrain from employing such data and instead use satellite-based data for all geo-location. Our motivation for this choice is twofold. First, since we are interested in the relative performance of LSTM compared to alternative forecasting methods, it is desirable to limit the effect of factors other than the forecasting method on empirical results. Ensuring consistency of the data sources helps with achieving this objective. Second, we aim at proposing a universally applicable framework for forecast model comparisons that facilitates testing models at any geo-location where satellite data is available.

Data was collected from the year 2005 to 2014 which is further divided into training, validation and test data sets. The test set, year 2014 in our analysis, is reserved for hold-out model comparisons. The validation set, year 2012 and 2013, is used to decide on model parameters (see Section 6). The rest of the data, i.e. years 2005–2011, is used for training the models.

5. Data Preparation

This section details activities in data preparation such as handling missing values and the input structure of the LSTM network in particular. With respect to missing values, we note that GEFS data exhibits ∼ 0.3% and ∼ 0.5% missing values for locations in the US and Europe, respectively. We impute these missing values for each atmospheric variable using a month-wise and step-wise linear regression fit. Of course, more powerful techniques such as, e.g., Kalman filters, could be considered and might facilitate a better imputation (Moritz et al., 2015; Soubdhan et al., 2016). On the other hand, the data is the same for all forecasting methods considered in the comparison. Consequently, the specific imputation strategy should not affect the relative performance of alternative forecasting methods.

Fig. 9 illustrates the input of the LSTM network. Recall that LSTM learns sequential data (see Section 3.1) and receives a total of 14 atmospheric input variables (see Table 2). Each GEFS input variable is a discrete time series of 7 steps per day. The blue[5] line exemplifies the distribution of an exemplary input variable throughout a day, with red dots representing variable values at different time steps. Values of individual time steps are sequentially sent to one single LSTM unit. The unit strives to understand the sequence. The connected LSTM units of Fig. 8 represent an unfolded view (in time) of a single LSTM unit, which is updating itself with every new input. The leftmost LSTM cell in Fig. 9 receives and processes information in the form of the 6 am value of a GEFS variable. The cell updates itself when it receives the next value of the 9 am time-step. At each time-step, LSTM training adjusts the model's understanding of the sequential behavior and relevance of the variable. Note that Fig. 9 simplifies the LSTM input structure in that it exemplifies the processing on one input variable. In our setup, all GEFS input variables are sent laterally to a LSTM unit for learning their interactive predictive ability. This architecture allows extracting information concerning the intrinsic dynamics of a time-series and how they are affected by external factors. More specifically, an LSTM unit combines the incoming signals and calculates an interaction of atmospheric variables at each time-step through its activation functions. The LSTM unit also receives the non-temporal variables (first three rows in Table 2) as input for each time-step.

In pre-tests on the AMS competition data, we also considered longer ranges of past data as network input. In particular, we evaluated input structures ranging from 1, 2, …, 30 previous days. Such tests revealed that LSTM performs best when receiving data from the previous day only as input. Therefore, we use this configuration for the forecasting

[5] For interpretation of color in Fig. 9, the reader is referred to the web version of this article.

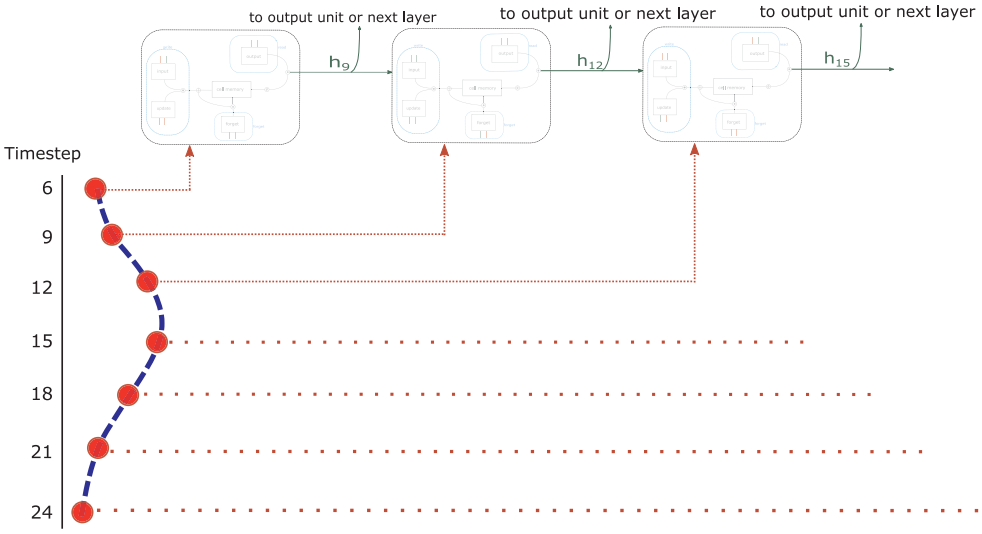


Fig. 9. Time ordered structure of input for feeding into LSTM units.

comparison to other methods and different geographical locations. The LSTM architecture is designed to handle sequential (temporal) data. This is not necessarily true for other prediction methods such as boosted trees. To ensure competitiveness of gradient boosting and other benchmarks, we refine the input structure according to standard practices (e.g., McGovern et al., 2015). In particular, the feature vector for benchmark methods includes the values of each variable across all time steps. As in the case of LSTM, the prediction target is the per-day observation of a given station. This approach has also been used by the winners of the AMS competition (McGovern et al., 2015).

6. Model tuning and benchmark algorithms

To put the performance of LSTM into context, we compare it to Gradient Boosting Regression (GBR), FFNN, and Persistence. FFNN has been the most popular approach for solar energy forecasting in previous research (see Table 1). Consequently, FFNN represents an important benchmark.

The motivation to consider Persistence is alike. Persistence assumes a scenario where there is no change in consecutive periodical measurements of readings. Hence, the Persistence model uses the measured target value at $t-1$ as prediction for time t. For day ahead prediction of GHI values, t is 24 h or one day and can be expressed via $GHI_{persistence,t} = GHI_{measured,t-1}$. Given that many previous studies employ Persistence, we considered it here in the interest of consistence with prior work.

With respect to GBR, we motivate its inclusion with the results of the AMS competition. A large set of diverse learning methods has been considered in this comparison, out of which GBR emerged as most powerful approach for solar energy forecasting (McGovern et al., 2015). In this sense, GBR can be considered the state of the art in the field, which suggests that comparisons to other - weaker - learning methods are neither necessary nor informative.

Hand (2006) raises the issue that researchers who introduce a novel prediction method may unduly (over-)tune this method to give it an advantage in empirical comparisons. Given that this is, to the best of our knowledge, the first evaluation of LSTM in solar energy prediction, we are determined to avoid such bias. To that end, we use the AMS competition as testbed and configure the LSTM architecture using the competition data. This provides a model skeleton, which we re-use in subsequent comparisons to GBR, FFNN, and Persistence. In particular, initial testing suggests an architecture of two layers of LSTM units followed by a drop-out layer and finally, a squashing sigmoid layer (see Section 8 for corresponding results).

The benchmark methods GBR and FFNN exhibit some meta-parameters that also require tuning. Furthermore, the data of the AMS competition is different from the GHI data that we use below to compare alternative forecasting models. With this in mind, we select one location from our GHI data collection and use it to fine-tune GBR, FFNN, and LSTM for the GHI data. In particular, we select one location in the USA, Penn State, for tuning. The Persistence model shows that GHI values for Penn State display the highest intermediate variability, which suggest that this location is particularly difficult to forecast. Hence, we select it for model tuning.

In particular, we use grid-search to identify suitable settings for algorithmic meta-parameters. The candidate parameter settings we consider are available in Table 4. The parameter values giving the best performance in the grid search stage (only Penn State) are then re-used for other location. More specifically, to grant forecasting models some freedom in adapting to a specific location, we fix all but one meta-parameter during grid search and tune the remaining free meta-parameter for each location individually. For LSTM and FFNN, the free meta-parameter is the number of iterations of the learning algorithm. For GBR, we tune the number of trees per location. The number of trees in a GBR model and the number of iterations in a neural network exhibit some similarity in the way they affect the learning algorithm. Both types of meta-parameters govern the degree to which an algorithm iteratively learns patterns in the training data. Higher (lower) values for these meta-parameters may thus result in over-(under-)fitting models. We argue that selecting conceptually similar meta-parameters for GBR, FFNN, and LSTM fine-tuning helps to ensure a fair comparison. To implement the neural network models LSTM and FFNN, we use the Python based Keras library (Chollet et al., 2015). GBR models are developed within in the R environment using the gbm package

Table 4
Meta-parameters used for grid search.

Algorithm	Meta-parameter	Values
LSTM	Neurons in 1st layer	c(10, 20, 40, 60, 80, 100, 110, 120, 140, 150)
	Neurons in 2nd layer	c(10, 20, 40, 60, 80, 100, 110, 120, 140, 150)
FFNN	Neurons in 1st layer	c(10, 20, 40, 60, 80, 100, 110, 120, 140, 150)
	Neurons in 2nd layer	c(10, 20, 40, 60, 80, 100, 110, 120, 140, 150)
GBR	Shrinkage	c(0.005, 0.01, 0.05, 0.1)
	Minobsinnode	c(8, 9, 10, 11, 12)
	Interaction depth	c(8, 9, 10, 11, 12)

(Ridgeway, 2015).

7. Performance metrics

We consider two performance metrics to assess forecasting accuracy: these are root mean square error (RMSE) and the mean absolute error (MAE) with their respective forecast skills. RMSE is more sensitive to large deviation between forecasts and actuals. MAE, on the other hand, is a suitable measure when assuming that forecast errors are proportional to energy costs. MAE is also the measure used in the AMS competition. RMSE and MAE are calculated by Eqs. (7) and (8), where N denotes the number of observations and I_x the value of the target variable with $x \in (predicted,measured)$. Forecast skill gives a relative measure of improvement in prediction over persistence model (Eq. (9)) where $ERR \in (RMSE,MAE)$.

$$RMSE = \frac{1}{\sqrt{N}}\sqrt{\sum_{i=1}^{N}(I_{(pred,i)}-I_{(meas,i)})} \tag{7}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|(I_{(pred,i)}-I_{(meas,i)})| \tag{8}$$

$$Forecast\ Skill = 1-\frac{ERR}{ERR_{persistence}} \tag{9}$$

8. Results

The following sections report results from initial experiments using AMS data and comparison of LSTM to benchmark methods across different locations in USA and Europe, respectively.

8.1. AMS competition

Experiments on the AMS data contrast alternative modeling regimes. The competition provides data for 14 years from 1994 to 2007. We use the years 1994–2002 for model training, the subsequent two years for validation, and the last two years for model testing.

To decide on the granularity of forecast model development, Table 5 compares the MAE of LSTM and GBR in a station-level as opposed to a region-level modeling approach. The former develops individual forecasting models for individual locations. The latter develops one model for the entire region of Oklahoma. Table 5 reveals a substantial improvement of the skill measure and corresponding reduction of MAE when adopting a regional modeling approach. This agrees with previous results of Lorenz et al. (2009a).

The regional modeling approach also reduces forecast variability. We measure forecast variability as the standard deviation of the forecasts for an individual station over the study period (i.e., test set). We obtain the standard deviation of the target variable (PV output values in the AMS competition) in the same way. Then, averaging over the absolute differences of the per-station forecast standard deviation and per-station target standard deviation, we find the mean absolute difference to be 395246.7 MJ/m^2 and 317256.3 MJ/m^2 for station-level and region-level forecasting, respectively. These numbers suggest the region-level modeling approach to reduce forecast variability by roughly 20%.

Table 5
Skill (%) and MAE (W m^{-2}) of region-level versus station-level forecast models for the AMS competition dataset.

Type	Forecast skill:MAE		MAE		
	GBR	LSTM	GBR	LSTM	Pers
Regional modeling	50.7	**51.3**	22.42	**22.16**	36.39
Per station modeling	38.4	**39.1**	23.31	**23.04**	47.32

Bold faces represent best prediction score.

Based on these results, we opt for a regional modeling approach and use this approach in subsequent experiments.

Finally, Table 5, we find strong evidence that LSTM outperforms GBR. To our knowledge, the test data of the AMS competition is not made available to the public. However, it is possible to make post-challenge submissions via the competition website. Creating such entry for the LSTM of Table 5 reveals that the MAE of LSTM is lower than that of the GBR-based winning entry of the AMS competition.

8.2. Experiments in other locations

Encouraging results on the AMS data motivate a broader evaluation of LSTM. Table 6 provides results in terms of RMSE and MAE across multiple locations. In the majority of cases, the best performance in both MAE and RMSE comes from the LSTM model. This provides strong evidence for LSTM being a very suitable technique for GHI prediction. More specifically, LSTM proves superior to GBR, FFNN, and Persistence. Considering the excellent performance of GBR in the AMS competition and the appealing performance of FFNN in much prior work on energy prediction, the superiority of LSTM as observed in Table 6 may generalize to other forecasting methods not considered here (i.e., weaker models compared to GBR and FFNN).

Although LSTM gives the most accurate GHI forecasts in most locations, Table 6 also indicates that our data set includes two locations, Palaiseau and Sonnblick, where LSTM forecasts notably less accurate than the benchmarks. To examine the cause of this behavior, we examine the development of RMSE across individual months in these two locations in Fig. 10(a) and (b). The analysis indicates that poor LSTM performance originates from inaccurate prediction in October. In other months, LSTM is competitive or better than the benchmark methods in both locations.

Fig. 10(a) and (b) put the impression that less accurate LSTM forecasts are characteristics to Palaiseau and Sonnblick into perspective and warrant further analysis whether the drop in forecast accuracy in October is location specific or general. To test this, we perform a month-wise analysis of LSTM versus GBR and FFNN. Results are available in Table 7, which depicts the percentage difference in RMSE between LSTM and GBR and FFNN, respectively. Positive values indicate that LSTM performs better (lower RMSE) than a benchmark. Table 7 also provides two monthly average percentage RMSE differences, which we calculate across all European and US regions, respectively.

Table 7 confirms weak performance in October to be a more general problem the LSTM. Across most regions in Europe, LSTM is inferior to the benchmarks. The average loss in RMSE compared to GBR and FFNN is 34.5% and 26.6% respectively. For almost every other month, LSTM improves upon GBR and FFNN in Europe. It is also noteworthy that we observe a few regions (e.g., Basel and Zurich), where LSTM October forecasts are better than those of FFNN and GBR. Yet, large forecast errors like those for Palaiseau and other European cities prevail. Corresponding results for US regions do not show abnormal results in October. Rather, the average (over US regions) performance of LSTM is better than that of the benchmarks in most months. Occasionally, LSTM predicts less accurately than GBR on average (e.g., January and February) but never worse than FFNN.

Further clarification of the origin of decreases in forecast accuracy is complicated because LSTM (as well as FFNN) constructs a highly nonlinear mapping between the input variables and the target variable. Although the internal mapping function is opaque, an analysis of the distribution of the input variables may shed some light on the source of forecast errors. For example, any forecasting model would suffer from an unusual distribution of input variables in the test set compared to the training data.

Table 8 provides average values for the GEFS based atmospheric input variables in October for the years 2013 and 2014. Recall that year 2013 belongs to the validation set whereas year 2014 represents the test

Table 6
Skill (%), RMSE ($W\,m^{-2}$) and MAE ($W\,m^{-2}$) of predicted GHI for chosen locations.

Country	Forecast skill:RMSE			RMSE				Forecast skill:MAE			MAE			
Location	FFNN	GBR	LSTM	FFNN	GBR	LSTM	Pers	FFNN	GBR	LSTM	FFNN	GBR	LSTM	Pers
Germany														
Hohenpeissenberg	54.1	54.5	**55.3**	30.03	29.79	**29.26**	65.48	53.4	54.2	**56.9**	21.28	20.91	**19.71**	45.69
Netherlands														
Cabauw	48.2	48.3	**53.1**	28.06	28.05	**25.42**	54.20	47.5	49.5	**51.0**	20.00	19.25	**18.67**	38.09
De Bilt	53.1	55.3	**55.4**	24.96	23.81	**23.76**	53.22	53.4	**55.1**	54.5	17.30	**16.66**	16.91	37.15
Austria														
Grossenzersdorf	**56.1**	49.9	55.8	**26.84**	30.67	27.04	61.19	52.4	47.7	**53.9**	20.24	22.20	**19.57**	42.47
Sonnblick	42.0	**44.2**	40.8	36.27	**34.89**	36.98	62.50	40.8	**42.9**	42.0	26.33	**25.39**	25.81	44.46
Wien	49.9	45.9	**55.6**	31.59	34.07	**27.94**	62.99	46.3	43.5	**53.5**	23.47	24.71	**20.34**	43.74
France														
Carpentras	57.4	59.2	**62.1**	26.87	25.70	**23.90**	63.05	55.2	55.6	**58.9**	18.85	18.71	**17.31**	42.11
Palaiseau	48.5	**53.5**	44.8	29.30	**26.46**	31.38	56.89	49.1	**50.9**	48.9	19.69	**19.00**	19.79	38.70
Spain														
Santander	53.7	53.0	**55.4**	31.43	31.93	**30.34**	67.95	54.4	53.9	**56.2**	22.45	22.68	**21.53**	49.20
La Coruna	44.0	47.6	**48.0**	33.34	31.21	**30.92**	59.50	43.6	48.1	**50.5**	24.18	22.24	**21.23**	42.87
Alicante	46.6	46.0	**46.9**	27.16	27.47	**26.98**	50.85	45.2	39.2	**47.2**	18.79	20.88	**18.11**	34.31
Madrid	49.5	**54.6**	53.1	25.43	**22.83**	23.60	50.31	42.1	50.9	**52.5**	19.12	16.22	**15.67**	33.02
Switzerland														
Basel	51.9	53.5	**58.0**	28.59	27.63	**24.99**	59.45	49.3	50.6	**55.5**	20.45	19.92	**17.95**	40.33
Zurich	50.2	53.7	**54.5**	30.61	28.46	**27.97**	61.52	43.8	49.5	**50.9**	23.00	20.70	**20.12**	40.96
Geneva	52.4	53.8	**53.8**	28.95	28.12	**28.07**	60.82	49.7	50.8	**52.7**	20.87	20.43	**19.64**	41.52
Davos	**52.2**	52.1	51.4	**32.12**	32.20	32.69	67.19	50.0	49.3	**50.7**	24.13	24.44	**23.78**	48.24
USA														
Desert Rock	37.6	42.0	**45.3**	29.38	27.29	**25.77**	47.07	40.7	**49.5**	48.6	17.58	**14.97**	15.24	29.63
Bondville	41.4	47.0	**49.5**	43.85	39.67	**37.78**	74.86	30.9	46.6	**48.2**	37.25	28.80	**27.94**	53.90
Fort Peck	37.7	39.7	**43.6**	37.00	35.82	**33.49**	59.36	35.2	40.7	**42.4**	28.18	25.81	**25.08**	43.50
Goodwin Creek	50.3	56.7	**57.9**	36.47	31.76	**30.89**	73.39	57.8	59.0	**59.2**	23.28	22.62	**22.49**	55.19
Penn State	50.0	53.0	**54.8**	38.71	36.39	**35.02**	77.48	48.6	54.2	**56.5**	29.53	26.31	**25.01**	57.44

Bold faces represent best prediction score.

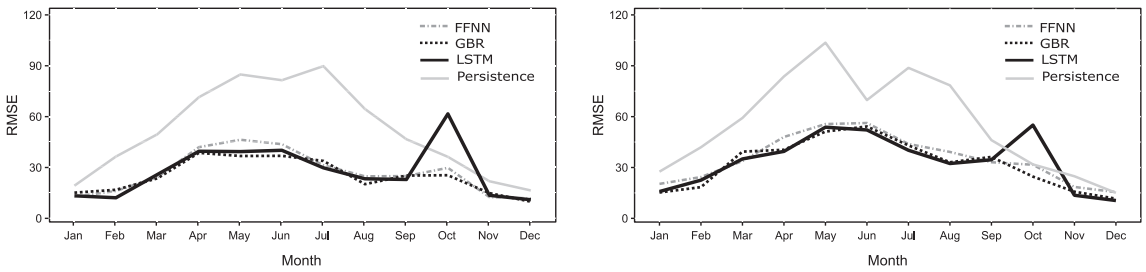
Pers abbreviates Persistence model.

set. This implies that 2013 has been used for model tuning whereas year 2014 has never been used during model development. Previous results including poor LSTM performance in October in Europe are based on the test set. Therefore, notable differences between the variable distributions between these month might deteriorate forecast accuracy. Table 8 shows results for two regions, Palaiseau and Zurich. LSTM October forecasts differs substantially across these regions. Performance in Palaiseau is especially poor whereas Zurich is one of the few examples where LSTMs outperforms the benchmarks.

To calculate the values of Table 8 we average over the 7 time-step values and 4 GEFS points surrounding the training region of the focal station (see Section 4.1.1 and 4.3). Table 8 also shows the yearly averages of input variable values for comparison.

Considering the input variable distributions in October, Table 8 shows some differences between the two years and locations, respectively. For example, considering the first input variable, 3-Hour accumulated precipitation at the surface, the difference between the means for 2013 and 2014 is negative for Palaiseau but positive for Zurich. Similarly, the difference between mean values for October 2013 and 2014 for variable two, downward long-wave radiative flux, is 0.9 for Palaiseau (320.7–319.8) and notably larger (4.7) for Zurich. Such differences in variable values across validation and test set months might have influenced LSTM performance. On the other hand, considering the corresponding differences for the yearly averages, we observe the same trend. For example, the differences between the yearly averages for 2013 and 2014 for the first input variable are -0.01 and 0.09 for Palaiseau and Zurich, respectively. One obtains almost exactly the same values when calculating the differences for October. We observe a similar pattern across all input variables, which suggest the variable distributions in October to be quite similar to that of other months in the year.

A different picture emerges when considering the differences between the standard deviations. Performing the same calculations as before (e.g., subtracting the standard deviation of a variable in 2013



(a) Palaiseau (b) Sonnblick

Fig. 10. Month-wise RMSE plots for worst performing locations of LSTM.

Table 7
Month wise improvement in RMSE score using LSTM over GBR and FFNN

Location	January		February		March		April		May		June		July		August		September		October		November		December	
	FFNN	GBR	FFNN	GBR	FFNN	GBR	FFNN	GBR	FFNN	GBR	FFNN	GBR	FFNN	GBR	FFNN	GBR	FFNN	GBR	FFNN	GBR	FFNN	GBR	FFNN	GBR
Europe																								
Hohenpeissenberg	20%	22%	18%	13%	23%	11%	8%	9%	2%	−10%	14%	3%	−17%	1%	7%	7%	2%	1%	−24%	−2%	7%	10%	20%	23%
Cabauw	15%	8%	13%	−3%	−13%	−8%	1%	11%	21%	8%	22%	19%	18%	5%	11%	21%	5%	29%	−33%	−45%	5%	−17%	13%	11%
De Bilt	8%	−12%	−2%	−12%	−8%	−16%	−5%	7%	11%	−3%	14%	5%	27%	3%	9%	14%	16%	16%	−92%	−35%	0%	−1%	−12%	1%
Grossenzersdorf	−7%	−18%	−11%	−2%	−19%	2%	10%	18%	−1%	6%	20%	16%	23%	36%	−6%	−18%	0%	19%	−25%	9%	8%	27%	23%	10%
Sonnblick	5%	−4%	−5%	−22%	−13%	11%	12%	2%	1%	−5%	13%	4%	8%	7%	8%	2%	−7%	5%	−77%	−123%	16%	13%	−2%	9%
Wien	14%	−1%	2%	8%	−34%	−5%	3%	21%	28%	24%	41%	26%	41%	40%	−10%	16%	13%	22%	−36%	1%	17%	32%	54%	31%
Carpentras	12%	−9%	11%	17%	19%	20%	9%	11%	9%	−2%	13%	−2%	−1%	16%	−14%	6%	−7%	−9%	28%	−12%	14%	29%	5%	12%
Palaiseau	5%	12%	16%	28%	−8%	−10%	10%	−2%	14%	−7%	6%	−9%	7%	12%	0%	−16%	17%	9%	−111%	−142%	1%	9%	−1%	−12%
Santander	4%	8%	−13%	0%	−13%	2%	−2%	−3%	22%	5%	−2%	10%	23%	24%	0%	2%	7%	16%	−21%	−50%	1%	0%	−11%	−5%
La Coruna	14%	25%	34%	13%	16%	−10%	−4%	3%	13%	0%	19%	6%	−11%	−2%	−5%	2%	8%	−2%	7%	−8%	8%	9%	−20%	−1%
Alicante	−5%	−7%	8%	−5%	2%	22%	1%	6%	−4%	−9%	0%	−12%	17%	13%	20%	28%	19%	21%	−15%	−29%	−5%	15%	−30%	12%
Madrid	−1%	9%	−6%	−4%	16%	20%	9%	21%	26%	3%	33%	7%	11%	19%	5%	−44%	11%	−13%	−28%	−89%	4%	−7%	−6%	11%
Basel	−1%	9%	−2%	1%	22%	−6%	16%	−4%	7%	19%	16%	3%	21%	12%	7%	21%	15%	−2%	13%	1%	−15%	16%	17%	21%
Zurich	15%	−5%	11%	−21%	19%	5%	−1%	−11%	−5%	−1%	12%	6%	4%	9%	10%	19%	15%	−13%	32%	8%	13%	−13%	−5%	−19%
Geneva	8%	−12%	0%	−18%	−2%	19%	−15%	8%	17%	1%	−9%	−4%	11%	4%	4%	4%	13%	19%	4%	15%	7%	−13%	−10%	−7%
Davos	3%	9%	7%	0%	24%	6%	3%	12%	2%	−1%	5%	−11%	−18%	−4%	6%	14%	10%	15%	−47%	−49%	12%	10%	5%	14%
Average	7%	2%	5%	−1%	2%	4%	3%	7%	10%	2%	14%	4%	10%	12%	3%	5%	9%	8%	−27%	−34%	6%	7%	2%	7%
US																								
Desert rock	−3%	25%	8%	3%	−4%	6%	−5%	−7%	−14%	3%	25%	26%	1%	0%	8%	19%	−35%	10%	26%	−10%	2%	−1%	−2%	1%
Bondville	3%	−2%	19%	−2%	38%	8%	6%	12%	16%	3%	33%	−5%	20%	−10%	40%	14%	42%	20%	32%	−2%	−5%	0%	5%	−2%
Fort peck	10%	−12%	6%	−5%	2%	7%	2%	4%	10%	−4%	25%	9%	10%	19%	15%	8%	12%	20%	32%	1%	8%	17%	−5%	0%
Goodwin creek	11%	−11%	28%	1%	−3%	0%	7%	19%	24%	−10%	−9%	−9%	−5%	−13%	11%	11%	−2%	10%	5%	14%	21%	18%	−4%	12%
Penn state	35%	−9%	−4%	−21%	14%	0%	−9%	7%	−15%	8%	21%	11%	33%	15%	16%	13%	8%	−1%	21%	−19%	1%	3%	23%	−1%
Average	11%	−2%	11%	−5%	9%	4%	0%	7%	4%	0%	19%	6%	12%	2%	18%	13%	5%	12%	23%	−3%	6%	7%	3%	2%

Table values denote the percentage improvement (positive sign) or reduction (negative sign) of RMSE of LSTM compared to FFNN and GBR, respectively. For example, a value of 20 % for Hohenpeissenberg in the January FFNN column indicates that the RMSE of LSTM is 20% better (lower) than that of FFNN.

We calculate averages RMSE reduction/improvement for each month and benchmark (FFNN or GBR) across all European and US regions, respectively.

Table 8
October and year-wise comparison of mean and standard deviation of atmospheric variables values corresponding to Palaiseau and Zurich for the year of 2013 and 2014.

Atmospheric input variable	Unit	Palaiseau								Zurich							
		October-2013		October-2014		Year-2013		Year-2014		October-2013		October-2014		Year-2013		Year-2014	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
3-Hour accumulated precipitation at the surface	$kg\,m^{-2}$	0.50	0.98	0.52	1.53	0.51	1.06	0.52	1.11	0.66	1.38	0.57	1.17	0.81	1.60	0.77	1.49
Downward long-wave radiative flux average at the surface	$W\,m^{-2}$	320.7	31.2	319.8	24.8	306.3	41.8	308.9	35.8	302.1	24.7	297.4	28.1	290.2	39.6	292.3	35.1
Downward short-wave radiative flux average at the surface	$W\,m^{-2}$	123.1	131.0	134.7	143.9	191.4	230.4	202.5	239.1	139.0	142.4	150.5	154.9	199.7	236.5	207.4	244.2
Air pressure at mean sea level	kPa	101.5	0.6	101.6	0.7	101.6	0.9	101.5	0.8	101.8	0.4	101.9	0.6	101.7	0.8	101.6	0.7
Precipitable water over the entire depth of the atmosphere	$kg\,m^{-2}$	21.3	6.7	21.4	5.7	17.5	8.0	18.3	7.4	16.4	5.0	16.1	5.2	13.8	6.8	14.5	6.2
Specific humidity at 2 m above ground	$kg\,kg^{-1}$	0.008	0.002	0.008	0.001	0.007	0.003	0.008	0.003	0.007	0.001	0.007	0.002	0.006	0.003	0.006	0.003
Total cloud cover over the entire depth of the atmosphere	%	0.110	0.157	0.113	0.181	0.097	0.156	0.084	0.156	0.131	0.180	0.095	0.162	0.130	0.202	0.117	0.189
Total column-integrated condensate over the entire atmos.	$kg\,m^{-2}$	0.111	0.156	0.097	0.196	0.098	0.156	0.084	0.154	0.131	0.180	0.095	0.161	0.130	0.201	0.117	0.188
Maximum temperature at 2 m above the ground	K	287.6	3.8	288.3	3.4	284.9	7.6	286.4	6.1	284.1	4.5	285.0	4.9	281.7	8.7	283.3	7.2
Minimum temperature hours at 2 m above the ground	K	285.0	3.9	285.5	3.3	281.9	7.1	283.4	5.7	280.5	4.0	280.7	4.6	277.8	8.1	279.3	6.5
Current temperature at 2 m above the ground	K	286.3	3.9	287.0	3.4	283.5	7.4	285.0	6.0	282.4	4.5	283.1	5.0	279.9	8.5	281.4	7.0
Upward long-wave radiation at the surface	$W\,m^{-2}$	380.7	21.9	384.0	21.1	367.5	39.9	374.6	33.8	361.1	22.4	363.9	26.0	349.5	44.4	356.8	37.7
Upward long-wave radiation at the top of the atmosphere	$W\,m^{-2}$	228.6	30.9	239.6	29.3	232.3	33.0	236.9	31.1	226.7	28.2	237.2	27.2	224.6	34.5	228.3	32.3
Upward short-wave radiation at the surface	$W\,m^{-2}$	26.9	28.9	29.9	31.7	40.9	46.1	42.1	46.9	27.9	28.9	30.9	31.8	40.6	44.7	39.4	43.9

from that in 2014 and comparing across locations), we observe some differences between results for October and the whole year. For several input variables (e.g., Maximum Temperature at 2 m above the ground and the subsequent four inputs in Table 8) the October differences are positive for Palaiseau but negative for Zurich. The situation changes when examining the standard deviations for the full years where results for the two locations are almost equivalent. For example, the difference between the standard deviation of Maximum Temperature at 2 m above the ground in 2013 (7.6) and 2014 (6.1) for Palaiseau is 1.5. We obtain exactly the same value for Zurich (8.7–7.2 = 1.5). Therefore, Table 8 provides some evidence that the variability of some input variables in October 2014 is different from previous periods. Learning intricate relationships between input variables and the target variable in the presence of unusual variation might have adversely affected LSTM predictions.

To obtain a clearer view on the distribution of input variables, we develop seasonal diagrams for a subset of the input variables. A seasonal diagram facilitates comparing variable values for the same periods (i.e., months) across multiple years (e.g. Crone and Kourentzes, 2010). Fig. 11 considers the years of the training, validation, and test period to clarify how test set forecasts of LSTM and other models might have been affected by irregular variable distributions. In the interest of brevity and comprehensibility, Fig. 11 considers only the eight most predictive variables, which we select in terms of GBR-based variable importance ranks (Friedman, 2001). The numbers on the y-axis indicate the rank of a variable in the GBR ranking, where a rank of one (eight) identifies the most (least) important variable.

Fig. 11 confirms the view that October 2014 displays some unusual variable developments. Several variables show regular seasonal patterns (numbering variables according to their importance ranks, this is true for variable 1, 2, 3, 5, 6) or the absence of seasonal effects (i.e., variable 8: air pressure at mean sea level). The remaining variables behave in a more erratic manner. Considering the third-most importance variable, we find substantial differences across the training, validation, and test years. Especially values from the test period differ substantially from those of preceding years. A sharp drop in September 2014 followed by an equally sharp increase in October may have had a particularly large effect on October forecasts; for example if forecasts embody a lagged influence of the corresponding input variable on the target variable. Variable 5 also shows a somewhat unusual behavior in 2014 in that September values are notably higher compared to the training and validation set. Again, such behavior will adversely affect models that capture a lag-effect between the variable and the target. Last, variable 7 (3-Hour accumulated precipitation at the surface), shows substantial variation in its monthly distribution across the training, validation, and test set. September values in the validation and test set evolve in opposite directions while both sets deviate largely from the training set in October. Clearly, such irregularity in a variable distribution across periods for model estimation, tuning, and application represent a big challenge for any forecasting model.

In summary, the seasonal diagrams of Fig. 11 provide strong evidence of abnormal behavior in several important input variables in the test period 2014. Fig. 12 further confirms this conclusion through analyzing daily averages of variable values in the month of October across training, validation, and test set years. The training set has the largest influence on a forecasting model since internal model parameters such as connection weights in a neural network are estimated from this data. The validation governs the selection of meta-parameters settings such as the number of hidden layers or hidden layer neurons in an neural network. Therefore, the validation set also affects forecast model development but to lesser extent than the training set. Fig. 12 shows how variable values differ substantially across periods in October; especially between the training and the test set.

In concluding the analysis of weak LSTM performance in one period and certain locations, we reiterate that LSTM as well as other powerful prediction methods are black-box procedures. Understanding their

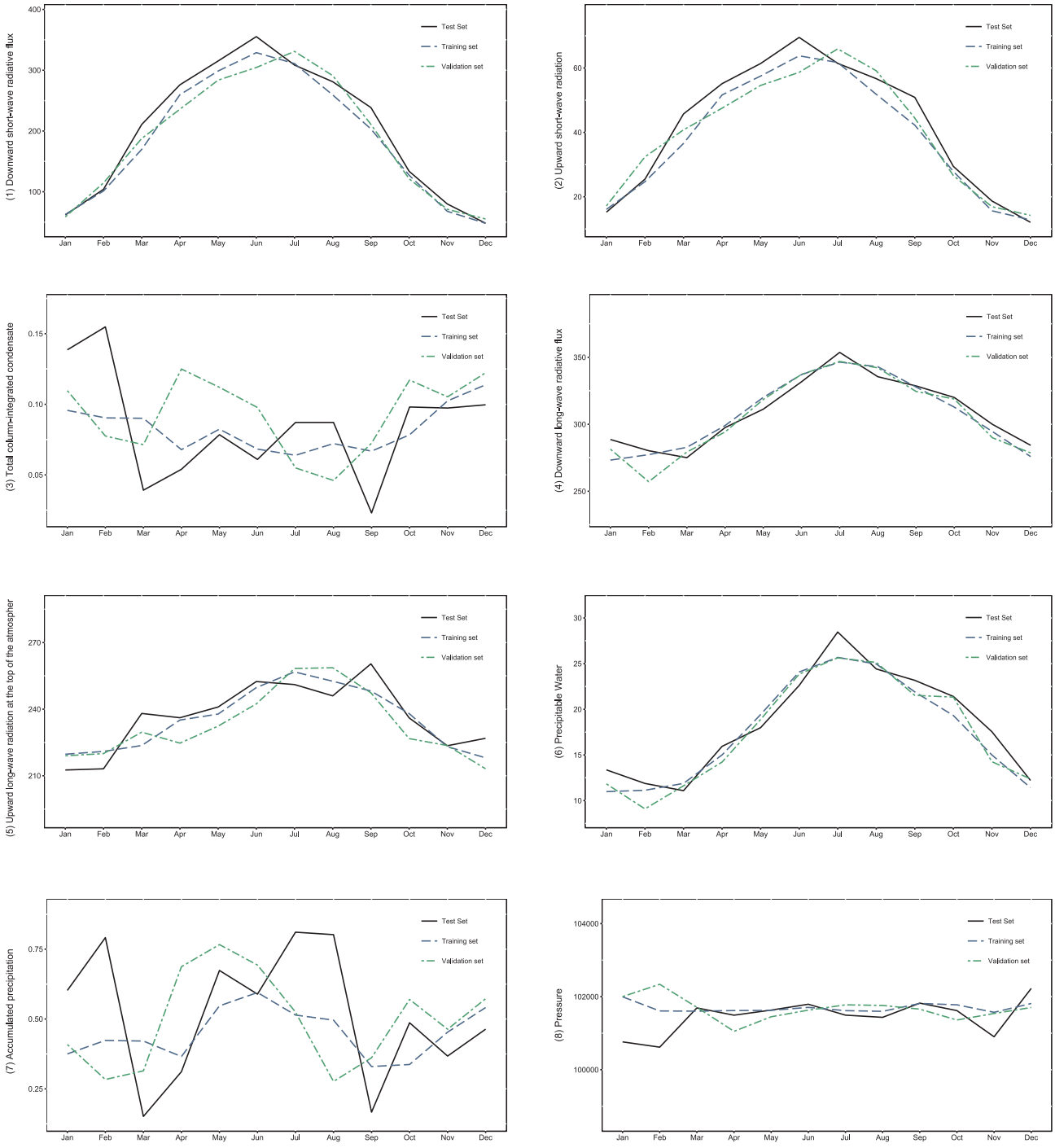


Fig. 11. Training, validation and test set comparison of top eight atmospheric input variables distribution in all month for Palaiseau (top eight variables from GBR model). Respective units are provided in Table 2.

internal mechanisms and interpreting the estimated relationship between input variables and the target is often impossible. This represent a drawback of such models compared to simpler linear forecasting methods such as dynamic linear regression or time-series models. Whether the gain in forecast accuracy justifies the loss of interpretability depends on the application of the model. Day-ahead forecasts of solar irradiance may support plant managers in their decision making processes, for example to facilitate an early participation in energy auction markets and efficient resource planning. We argue that forecast accuracy is essential in such context since the economic consequences of inappropriate buying/selling/hedging decisions may be severe. Under this assumption, the use of nonlinear methods is justified, although such methods are opaque.

The empirical comparison suggests that LSTM forecasts are more accurate than forecasts of the FFNN, GBR, and Persistence benchmark on average. The follow-up analysis of situations where the benchmarks beat LSTM has provided evidence that the distribution of several input variables displays irregularity in these situations. Deviations between training versus test set variable distributions affect every forecasting model; not only LSTM. However, the overall superiority of LSTM must come from it succeeding in distilling more predictive information from the input variables than the benchmarks. Therefore, distortion of the distribution of some input variables is likely to hurt LSTM more severely than benchmark models. This could explain the relatively poor

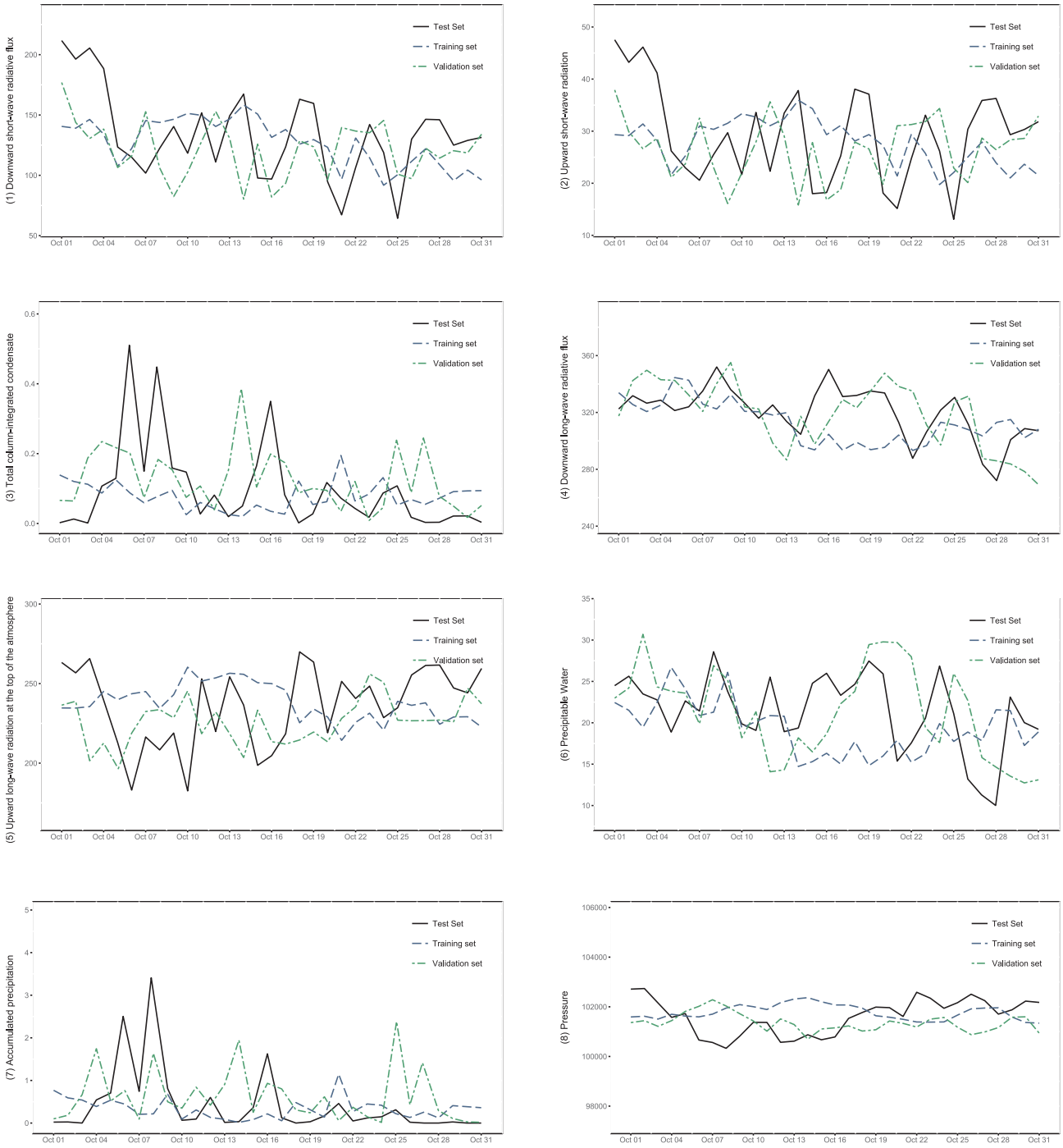


Fig. 12. Training, validation and test set comparison of top eight atmospheric input variables distribution in October month for Palaiseau (top eight variables from GBR model). Respective units are provided in Table 2.

results for LSTM in October, which we observe in Table 7.

However, Table 7 also shows that not only LSTM face difficulties in certain months/regions. For example, FFNN performs relatively poor in June. GBR also exhibits weaknesses in certain periods. For example, July performance of GBR in Europe appears relatively poor in that GBR performs less - and typically substantially less - good than LSTM in all but two European locations. In this sense, the empirical results indicate that temporal decreases in forecast accuracy are not specific to LSTM but also occur with other forecasting methods; at least the methods considered here. Model combination approaches might be a way forward. One could envision an ensemble approach that switches automatically from one model to another in response to signals from region/period indicators. Similar concepts have been advanced in the scope of mixture-of-expert approaches (Yuksel et al., 2012). Evaluating related techniques in GHI and more generally energy forecasting seems to be a promising avenue for future research. As far as this experiment is concerned, overall forecasting results across all regions and periods provide strong evidence in favor of LSTM. It is the strongest method in the comparison and predicts more accurately than FFNN and GBR on average.

9. Conclusions

The paper studies LSTM as a powerful approach for time series

forecasting and examines its potential to predict solar irradiance. We observe LSTM to outperform challenging machine learning benchmarks, which evidences its ability to forecast irradiance with high accuracy.

The contributions of the paper are threefold. First, the study demonstrates the robustness of LSTM and suggests that a properly configured LSTM model outperforms GBR and FFNN for day-ahead GHI predictions. Therefore, forecasters are well-advised to consider LSTM when using machine learning methods in irradiance prediction.

The experimental framework represents the second contribution. Given the difficulties associated with gathering representative data, many previous studies compared alternative forecasting methods at a single location. However, location specific results require external validation at other regions with different spatial and climatic conditions. Approximation of GHI ground readings using satellite based solar models and development of region-based forecast models through creating a neighborhood of virtual stations overcomes this limitation.

Combining the experimental framework and LSTM, a third contribution emerges in the form of enhanced remote solar resource assessment. For resource assessment, planning authorities usually look at the long-range horizon availability of solar energy in the location (Sengupta et al., 2015). Even though a location might provide annually high solar energy, for PV setup, as an initial measure, the short-horizon predictability must be taken into account for efficient early participation in energy auction market (e.g., when the respective PV system goes live). As shown here, the intermittent nature of solar resources varies over different places. The Persistence score, by definition, explains day-ahead variability in GHI of a location. The correlation of RMSE scores between Persistence and other ML models is ~ 0.6. That is, an increase in the Persistence error does not necessarily imply a corresponding increase in the forecast errors of ML algorithms. Rather, ML models have been shown to understand atmospheric behavior even in cases of high-variability. For example, a retrospective check using the experimental design proposed here, which allows remote modeling of any location where accurate now-casted GHI data is available, will inform decision-making regarding location choice.

In view of the encouraging results observed in the study, one might wonder whether data-driven learning methods can eventually replace the more complicated NWP models, which prevail in energy modeling. In our opinion, learning methods are powerful tools and deserve a place in solar energy modeling. However, they also suffer limitations. The black-box character of learning algorithms prohibits understanding and verifying the intricate relationships between variables that a learning model embodies. In view of this, a co-existence between data-driven and NWP models seems a more likely scenario. In fact, a promising avenue for future research could be to examine hybrid models that combine physical and learning models in a best-of-breed approach. For example, interpretable - say linear - models often display high bias because they lack the expressive power to capture intricate (non-linear) patterns in a time series. In a hybrid forecasting framework, one could augment the linear component with a black-box machine learning algorithm the task of which could be to model systematic patterns in the residuals of the linear model. Such combination may achieve a favorable balance between model interpretability and forecasting accuracy. Corresponding two-model approaches have been studied in the time series forecasting literature (e.g. Chen and Wang, 2007). It seems fruitful to investigate comparable hybrid modeling schemes of NWP and data-driven models in solar energy forecasting.

The present study also exhibits several limitations, which open paths for future research. For example, many solar energy prediction models use outputs of mesoscale NWP models. Therefore, a comprehensive comparison of state of the art NWP to advanced ML models such as LSTM would be valuable to better appraise the benefits of potentially better accuracy of NWP models and the costs associated with developing corresponding forecasts. Furthermore, LSTM is just one approach of the wide family of deep learning neural networks. Alternative techniques such as convolutional neural networks or Bayesian deep learning might also have potential for energy modeling. For example, convolutional neural nets could be effective for spatial averaging using weather variables as a filter to learn geographical regions. In general, the expanding universe of predictive methods should be thoroughly integrated with the solar energy domain to advance their economic feasibility, efficient market penetration and eventually pave the way for unlocking the biggest source of energy.

Acknowledgement

We would like to thank the managing editor, Prof. Hugo Pedro for all his efforts in handling the paper. We are particularly grateful to two anonymous reviewers who provided very helpful comments how to improve the paper. Likewise, we very much appreciate the support of Prof. Jan Kleissl whose comments on an initial version of the paper have been extremely helpful.

References

Aler, R., Martín, R., Valls, J.M., Galván, I.M., 2015. A study of machine learning techniques for daily solar energy forecasting using numerical weather models. In: Intelligent Distributed Computing VIII. Springer, pp. 269–278. <http://link.springer.com/10.1007/978-3-319-10422-5_29>.

Armstrong, J.S., Fildes, R., 2006. Making progress in forecasting. Int. J. Forecast. 22 (3), 433–441. <http://www.sciencedirect.com/science/article/pii/S0169207006000513>.

Bacher, P., Madsen, H., Nielsen, H.A., 2009. Online short-term solar power forecasting. Solar Energy 83 (10), 1772–1783. http://dx.doi.org/10.1016/j.solener.2009.05.016.

Banfield, R.E., Hall, L.O., Bowyer, K.W., Kegelmeyer, W.P., 2005. Ensemble diversity measures and their application to thinning. Inf. Fus. 6 (1), 49–62.

Bengio, Y., Simard, P., Frasconi, P., 1994. Learning long-term dependencies with gradient descent is difficult. IEEE Trans. Neural Netw. 5 (2), 157–166. <http://ieeexplore.ieee.org/document/279181/>.

Brancucci Martinez-Anido, C., Botor, B., Florita, A.R., Draxl, C., Lu, S., Hamann, H.F., Hodge, B.M., 2016. The value of day-ahead solar power forecasting improvement. Solar Energy 129, 192–203. http://dx.doi.org/10.1016/j.solener.2016.01.049.

Cao, J., Lin, X., 2008. Study of hourly and daily solar irradiation forecast using diagonal recurrent wavelet neural networks. Energy Convers. Manage. 49 (6), 1396–1406. <http://linkinghub.elsevier.com/retrieve/pii/S0196890408000125>.

Caruana, R., Niculescu-Mizil, A., 2006. An empirical comparison of supervised learning algorithms. In: Proceedings of the 23rd International Conference on Machine Learning. ACM, pp. 161–168.

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 (11), 2856–2870. http://dx.doi.org/10.1016/j.solener.2011.08.027.

Chen, K.-Y., Wang, C.-H., 2007. A hybrid Sarima and support vector machines in forecasting the production values of the machinery industry in taiwan. Expert Syst. Appl. 32 (1), 254–264.

Chollet, F., et al., 2015. Keras < https://github.com/fchollet/keras > .

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 (11), 2881–2893. <http://www.sciencedirect.com/science/article/pii/S0038092X11002982>.

Cococcioni, M., D'Andrea, E., Lazzerini, B., 2011. 24-hour-ahead forecasting of energy production in solar PV systems. In: 2011 11th International Conference on Intelligent Systems Design and Applications. No. November. IEEE, pp. 1276–1281. <http://ieeexplore.ieee.org/document/6121835/>.

Cornaro, C., Pierro, M., Bucci, F., 2015. Master optimization process based on neural networks ensemble for 24-h solar irradiance forecast. Solar Energy 111, 297–312. http://dx.doi.org/10.1016/j.solener.2014.10.036.

Crone, S.F., Kourentzes, N., 2010. Feature selection for time series prediction a combined filter and wrapper approach for neural networks. Neurocomputing 73 (10), 1923–1936.

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.

Ding, M., Wang, L., Bi, R., 2011. An ANN-based approach for forecasting the power output of photovoltaic system. Proc. Environ. Sci. 11 (PART C), 1308–1315. http://dx.doi.org/10.1016/j.proenv.2011.12.196.

Duffy, A., Rogers, M., Ayompe, L., 2015. Renewable Energy and Energy Efficiency-Assessment of Projects and Policies. Wiley-Blackwell. <http://eu.wiley.com/WileyCDA/WileyTitle/productCd-1118631048,subjectCd-ECD0.html>.

Ekici, B.B., 2014. A least squares support vector machine model for prediction of the next day solar insolation for effective use of PV systems. Measurement 50, 255–262. <http://www.sciencedirect.com/science/article/pii/S0263224114000165>.

Fernández, Á., Gala, Y., Dorronsoro, J.R., 2014. Machine learning prediction of large area photovoltaic energy production. In: Data Analytics for Renewable Energy Integration. Springer, pp. 38–53. <http://link.springer.com/10.1007/978-3-319-13290-7_3>.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Statist. 29 (5), 1189–1232.

Gala, Y., Fernández, Á., Díaz, J., Dorronsoro, J.R., 2016. Hybrid machine learning forecasting of solar radiation values. Neurocomputing 176, 48–59. http://dx.doi.org/10.

1016/j.neucom.2015.02.078.
Graves, A., 2012. Supervised Sequence Labelling. Springer, pp. 5–13. <http://link.springer.com/10.1007/978-3-642-24797-2_2>.
Hamill, T., Bates, G., Whitaker, J., Murray, D., Fiorino, M., Galarneau, T., 2013. A Description of the 2nd Generation NOAA Global Ensemble Reforecast Data Set. Tech. rep < https://www.esrl.noaa.gov/psd/forecasts/reforecast2/README.GEFS_Reforecast2.pdf > .
Hand, D.J., 2006. Classifier technology and the illusion of progress. Statist. Sci. 21 (1), 1–14. <http://projecteuclid.org/Dienst/getRecord?id=euclid.ss/1149600839/>.
Hochreiter, S., Bengio, Y., Frasconi, P., Schmidhuber, J., 2001. Gradient Flow in Recurrent Nets: The Difficulty of Learning Long-Term Dependencies < http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.24.7321 > .
Inage, S.-i., 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://www.sciencedirect.com/science/article/pii/S0038092X17303961>.
Inman, R.H., Pedro, H.T.C., Coimbra, C.F.M., 2013. Solar forecasting methods for renewable energy integration. Prog. Energy Combust. Sci. 39 (6), 535–576. http://dx.doi.org/10.1016/j.pecs.2013.06.002.
Junior, J.G. d. S.F., Oozeki, T., Ohtake, H., Shimose, K.-i., Takashima, T., Ogimoto, K., 2014. Regional forecasts and smoothing effect of photovoltaic power generation in Japan: An approach with principal component analysis. Renew. Energy 68 (August), 403–413. <http://linkinghub.elsevier.com/retrieve/pii/S0960148114000986>.
Kemmoku, Y., Orita, S., Nakagawa, S., Sakakibara, T., 1999. Daily insolation forecasting using a multi-stage neural network. Solar Energy 66 (3), 193–199. <http://linkinghub.elsevier.com/retrieve/pii/S0038092X99000171>.
Kleissl, J., 2013. Solar Energy Forecasting and Resource Assessment. Academic Press. <http://www.sciencedirect.com/science/article/pii/B9780123971777000152>.
Kostylev, V., Pavlovski, A., 2011. Solar Power Forecasting Performance - Towards Industry Standards. In: 1st International Workshop on the Integration of Solar Power into Power Systems, Aarhus, Denmark < https://ams.confex.com/ams/92Annual/webprogram/Manuscript/Paper203131/AMS_VK_AP_Paper2011submitted.pdf > .
Larson, V.E., 2013. Forecasting solar irradiance with numerical weather prediction models. In: Solar Energy Forecasting and Resource Assessment. Elsevier, pp. 299–318. http://dx.doi.org/10.1016/B978-0-12-397177-7.00012-7.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444. https://doi.org/10.1038/nature14539.
Leva, S., Dolara, A., Grimaccia, F., Mussetta, M., Ogliari, E., 2017. Analysis and validation of 24 hours ahead neural network forecasting of photovoltaic output power. Math. Comput. Simul. 131, 88–100. http://dx.doi.org/10.1016/j.matcom.2015.05.010.
Li, Z., Rahman, S., Vega, R., Dong, B., 2016. A hierarchical approach using machine learning methods in solar photovoltaic energy production forecasting. Energies 9 (1), 55. <http://www.mdpi.com/1996-1073/9/1/55>.
Linares-Rodríguez, A., Ruiz-Arias, J.A., Pozo-Vázquez, D., Tovar-Pescador, J., 2011. Generation of synthetic daily global solar radiation data based on ERA-Interim reanalysis and artificial neural networks. Energy 36 (8), 5356–5365. http://dx.doi.org/10.1016/j.energy.2011.06.044.
Long, H., Zhang, Z., Su, Y., 2014. Analysis of daily solar power prediction with data-driven approaches. Appl. Energy 126, 29–37. http://dx.doi.org/10.1016/j.apenergy.2014.03.084.
Lorenz, E., Heinemann, D., 2012. Prediction of Solar Irradiance and Photovoltaic Power, vol. 1 Elsevier Ltd.
Lorenz, E., Hurka, J., Heinemann, D., Beyer, H.G., 2009a. Irradiance forecasting for the power prediction of grid-connected photovoltaic systems. IEEE J. Select. Top. Appl. Earth Observ. Rem. Sens. 2 (1), 2–10. <http://ieeexplore.ieee.org/document/4897348/>.
Lorenz, E., Remund, J., Müller, S.C., Traunmüller, W., Steinmaurer, G., Pozo, D., Antonio, J., Fanego, V.L., Ramirez, L., Romeo, M.G., Kurz, C., Pomares, L.M., Guerrero, C.G., 2009b. Benchmarking of different approaches to forecast solar irradiance. In: Proceedings of the 24th European Photovoltaic Solar Energy Conference, Hamburg, Germany. Hamburg, pp. 4199–4208 < http://archive.iea-shc.org/publications/downloads/24th_EU_PVSEC_5BV.2.50_lorenz_final.pdf > .
Lorenz, E., Scheidsteger, T., Hurja, J., Heinemann, D., Kurz, C., 2010. Regional PV power prediction for improved grid integration. Prog. Photovolt: Res. Appl. 19 (2011), 757–771. http://dx.doi.org/10.1002/pip.1033.
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://www.sciencedirect.com/science/article/pii/S0038092X15005897>.
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 (5), 746–756. http://dx.doi.org/10.1016/j.solener.2011.01.007.
Mathiesen, P., Kleissl, J., 2011. Evaluation of numerical weather prediction for intra-day solar forecasting in the continental united states. Solar Energy 85 (5), 967–977. <http://www.sciencedirect.com/science/article/pii/S0038092X11000570>.
McGovern, A., Gagne, D.J., Basara, J., Hamill, T.M., Margolin, D., 2015. Solar energy prediction: an international contest to initiate interdisciplinary research on compelling meteorological problems. Bull. Am. Meteorol. Soc. 96 (8), 1388–1393. http://dx.doi.org/10.1175/BAMS-D-14-00006.1.
Moreno, A., Gilabert, M.A., Martínez, B., 2011. Mapping daily global solar irradiation over Spain: a comparative study of selected approaches. Solar Energy 85 (9), 2072–2084. <http://www.sciencedirect.com/science/article/pii/S0038092X11001976>.
Moritz, S., Sard, A., Bartz-Beielstein, T., Zaefferer, M., Stork, J., 2015. Comparison of Different Methods for Univariate Time Series Imputation in r. CoRR abs/1510.03924 < http://arxiv.org/abs/1510.03924 > .
Nanculef, R., Valle, C., Allende, H., Moraga, C., 2006. Ensemble learning with local diversity. Artif. Neural Netw.–ICANN 2006, 264–273.
Nottrott, A., Kleissl, J., 2010. Validation of the NSRDB–SUNY global horizontal irradiance in California. Solar Energy 84 (10), 1816–1827.
Paoli, C., Voyant, C., Muselli, M., Nivet, M.L., 2010. Forecasting of preprocessed daily solar radiation time series using neural networks. Solar Energy 84 (12), 2146–2160. http://dx.doi.org/10.1016/j.solener.2010.08.011.
Peel, M.C., Finlayson, B.L., McMahon, T.A., 2007. Updated world map of the Köppen-Geiger climate classification. Hydrol. Earth Syst. Sci. Disc. 4 (2), 439–473. <https://hal.archives-ouvertes.fr/hal-00298818/>.
Perez, R., Lorenz, E., Pelland, S., Beauharnois, M., Van Knowe, G., Hemker, K., Heinemann, D., Remund, J., Müller, S.C., Traunmüller, W., Steinmauer, G., Pozo, D., Ruiz-Arias, J.A., Lara-Fanego, V., Ramirez-Santigosa, L., Gaston-Romero, M., Pomares, L.M., 2013. Comparison of numerical weather prediction solar irradiance forecasts in the US, Canada and Europe. Solar Energy 94 (August), 305–326. <http://linkinghub.elsevier.com/retrieve/pii/S0038092X13001886>.
Perez, R., Zweibel, K., Hoff, T.E., 2011. Solar power generation in the US: Too expensive, or a bargain? Energy Policy 39 (11), 7290–7297. http://dx.doi.org/10.1016/j.enpol.2011.08.052.
Podestá, G.P., Núñez, L., Villanueva, C.A., Skansi, M.A., 2004. Estimating daily solar radiation in the Argentine Pampas. Agricult. For. Meteorol. 123 (1–2), 41–53. <https://miami.pure.elsevier.com/en/publications/estimating-daily-solar-radiation-in-the-argentine-pampas>.
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.
Ridgeway, G., 2015. gbm: Generalized Boosted Regression Models. R package version 2.1.1. <https://CRAN.R-project.org/package=gbm>.
Robusto, C.C., 1957. The cosine-haversine formula. Am. Math. Mont. 64 (1), 38–40.
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.
Sangster, A.J., 2014. Solar Photovoltaics. In: 8th PVSEC. No. September, pp. 145–172 < http://link.springer.com/10.1007/978-3-319-08512-8_7 > .
Schmidhuber, J., 2015. Deep learning in neural networks: an overview. Neural Netw. 61, 85–117.
Schroedter-Homscheidt, M., 2016. The Copernicus Atmosphere Monitoring Service (CAMS) Radiation Service in a Nutshell. Tech. rep.
Sengupta, M., Habte, A., Kurtz, S., Dobos, A., Wilbert, S., Lorenz, E., Stoffel, T., Renné, D., Gueymard, C., Myers, D., Wilcox, S., Blanc, P., Perez, R., 2015. Best Practices Handbook for the Collection and Use of Solar Resource Data for Solar Energy Applications Best Practices Handbook for the Collection and Use of Solar Resource Data for Solar Energy Applications. Technical Report, NREL/TP-5D00-63112 (February), pp. 1–255.
Sengupta, M., Weekley, A., Habte, A., Lopez, A., Molling, C., 2012. Validation of the National Solar Radiation Database (NSRDB) Preprint. Nrel (September) < http://www.nrel.gov/docs/fy15osti/64981.pdf > .
Soman, S.S., Zareipour, H., Malik, O., Mandal, P., 2010. A review of wind power and wind speed forecasting methods with different time horizons. In: North American Power Symposium 2010, NAPS 2010 (October) < http://ieeexplore.ieee.org/document/5619586/ > .
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://www.sciencedirect.com/science/article/pii/S0038092X16001444>.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15 (April), 1929–1958. <http://jmlr.org/papers/v15/srivastava14a.html>.
Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9.
Tadesse, A., Kankiewicz, A., Perez, R., Lauret, P., 2016. Day ahead irradiance forecast variability characterization using satellite data. In: 43rd IEEE Photovoltaic Specialists Conference.
Voyant, C., Notton, G., Kalogirou, S., Nivet, M.-L., Paoli, C., Motte, F., Fouilloy, A., 2017. 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.
Voyant, C., Randimbivololona, P., Nivet, M.L., Paoli, C., Muselli, M., 2014. Twenty four hours ahead global irradiation forecasting using multi-layer perceptron. Meteorol. Appl. 21 (3), 644–655. http://doi.wiley.com/10.1002/met.1387.
Wang, G., Su, Y., Shu, L., 2016. One-day-ahead daily power forecasting of photovoltaic systems based on partial functional linear regression models. Renew. Energy 96, 469–478. http://dx.doi.org/10.1016/j.renene.2016.04.089.
Wang, X., Guo, P., Huang, X., 2011. A review of wind power forecasting models. Energy Proc. 12, 770–778. the Proceedings of International Conference on Smart Grid and Clean Energy Technologies (ICSGCE 2011). <http://www.sciencedirect.com/science/article/pii/S1876610211019291>.
Yona, A., Senjyu, T., Saber, A.Y., Funabashi, T., Sekine, H., Kim, C.-H., 2007. Application of neural network to one-day-ahead 24 hours generating power forecasting for photovoltaic system. In: 2007 International Conference on Intelligent Systems Applications to Power Systems, November. IEEE, pp. 1–6. <http://ieeexplore.ieee.org/document/4441657/>.
Yuksel, S.E., Wilson, J.N., Gader, P.D., 2012. Twenty years of mixture of experts. IEEE Trans. Neural Netw. Learn. Syst. 23 (8), 1177–1193.