Objective Framework for Optimal Distribution of Solar Irradiance Monitoring Networks

Athanassios Zagouras^a, Alexander Kolovos^b, Carlos F.M. Coimbra^{a,*}

^a*Department of Mechanical and Aerospace Engineering, Jacobs School of Engineering
Center of Excellence in Renewable Energy Integration and Center for Energy Research
University of California San Diego La Jolla, California, 92093, USA*

^b*SpaceTimeWorks LLC, San Diego, CA, USA*

Abstract

Time-resolved characterization of solar irradiance at the ground level is a critical element in solar energy analysis. Siting of nodes in a network of solar irradiance monitoring stations (MS) is a multi-faceted problem that directly affects the determination of the solar resource and its spatio-temporal variability. The present work proposes an objective framework to optimize the deployment of solar MS over a sub-continental region. There are two main components in the proposed methodology. The first employs cluster analysis using the affinity propagation algorithm, to select the optimal number of clusters (regions with coherent solar microclimates) upon internal coherence criteria. The second component employs stochastic prediction and validation, through the use of a Bayesian maximum entropy method, and selects the optimal MS configuration, according to geostatistical criteria, among the solutions recommended by the cluster analysis. We apply this two-pronged methodology to determine clusters and optimal locations for global horizontal irradiance monitoring across the state of California. In this proof-of-concept study, 3 disparate MS configurations are examined within the cluster partition. The subsequent geostatistical analysis indicates that all configurations rank almost equally well based on different statistical error measures. The optimal configuration can be singled out depending on desired criteria of choice.

Keywords: solar data, cluster analysis, adaptive affinity propagation, Bayesian maximum entropy, spatiotemporal prediction, geostatistics

1. Introduction

Ground-based solar irradiance monitoring stations (MS) are the most reliable source to provide accurate assessment of ground solar radiation. Ground-based monitoring networks have been installed throughout the world for different purposes, such as the Baseline Surface Radiation Network [1], the European Solar Radiation Atlas [2], the California Irrigation Management Information System (CIMIS). Common limitations in such networks are (i) the limited spatial coverage compared to satellite modeled irradiance data in high spatial resolution gridded domains, and (ii) the lack of prior knowledge regarding the MS spatial representativeness that could enable the development of operational plans about MS installation locations. Traditional network design analysis relies mainly on topological design to determine locations for MS.

In recent years, advanced data mining methods (e.g., cluster analysis) have provided insight on the coherence of spatial groups. This feature has helped clustering methods find application in designing monitoring station networks; see, e.g., [3]. At the same time, such methods often require a relatively large number of observations, which typically

*Corresponding author

Email addresses: azagouras@eng.ucsd.edu (Athanassios Zagouras), alexander.kolovos@spacetimeworks.com (Alexander Kolovos), ccoimbra@ucsd.edu (Carlos F.M. Coimbra)

exceeds the number of monitoring stations within a region. Geophysical clustering methods can be enhanced substantially by remote sensing, which provides extensive attribute coverage in increasingly detailed spatial and temporal resolutions. A shortcoming of satellite data is the relatively low accuracy and spatio-temporal resolution levels. Most radiation models from remote sensing depend strongly on calibration sites and dynamic atmospheric parameters; for example, satellite data of global horizontal irradiance (GHI) and direct normal irradiance (DNI) incur errors of the order of 5–40% depending on location and time granularity [4, 5].

Several studies assess the representativeness of measuring points by comparing solar observations measured from both satellites and ground sites [6, 7, 8]. In related studies, ground-based stations are members of existing sensor networks and they are used to define the *ground-truth*. Zagouras *et al.* [9, 10, 11] were the first to use cluster analysis of features extracted from satellite-derived solar data to determine the appropriate number of coherent clusters within different domains of interest. The results of these studies were based on criteria related to the convergence of the clustering methods to cluster partitions, validated with clustering validity metrics. An appropriate number of clusters is determined by assessing the clustering quality among a variety of clustering solutions.

In addition to clustering, geostatistics is also used in network design studies. For example, in the presence of sparse measurements, [12] show how geostatistical predictive techniques in conjunction with simulation iterations can assist in determining optimized locations. More commonly, though, geostatistics is utilized to model an attribute in space and/or time from a set of measured values. Specifically, in solar literature solar irradiance is often modeled by using parametric models, e.g., [13], autoregressive techniques for solar irradiance time series, e.g., [14], and geostatistics. Geostatistics has the benefit of considering solar irradiance as a stochastic random field [15], thus providing insight about the solar irradiance field structure, homogeneities/stationarity and internal characteristics. In this process, geostatistics enables integration of relevant information about the solar irradiance field, and utterly enables prediction on the basis of spatial and/or temporal correlation.

Most often, methodologies used for geostatistical prediction are linear model-based techniques such as the kriging family of predictors. Kriging techniques have been previously used to predict solar irradiance at unsampled locations, both within purely spatial and spatiotemporal contexts (e.g., [16, 17, 18]). More recently, [19] exhibited a more in-depth kriging example in the joint spatiotemporal continuum, and indicated the importance of spatiotemporal prediction for solar irradiance forecasting. Despite their mainstream character, these geostatistical techniques are built on restrictive assumptions and known limitations. For example, linear models are used to describe phenomena that are inherently nonlinear, and Gaussian assumptions are required for the data distributions by the linear interpolators. An additional major weakness of mainstream techniques is the inability to account rigorously for nontrivial uncertainty in the data set. Thus, in the presence of uncertain measurements such as probabilistic distributions or interval data, measurements are either skipped or being used by reducing their informational content to single values.

We propose the Bayesian maximum entropy (BME) method as an advanced alternative to the mainstream geostatistical methodologies. Being free of limitations and weaknesses like the above, BME additionally features very attractive characteristics for solar irradiance studies, such as the ability to rigorously incorporate uncertain data. This methodology has been applied previously broadly and successfully in fields such as environmental sciences, atmospheric monitoring, and environmental health risk and assessment; see, e.g., [20, 21, 22, 23]. BME as a valuable alternative for solar irradiance prediction was first introduced by Kolovos in [24] in a limited study across the USA. An extensive application of BME in accurately predicting subhourly solar photovoltaic output over state-size areas was presented more recently by Lee *et al.* in [25]. On the side of teaming with cluster analysis in the proposed framework, our study extends the previous geostatistical solar literature by applying BME prediction for solar irradiance on a spatially large scale in the state-wide domain of California, and a temporally systematic scale at subhourly 30-minute intervals over a period of days across a calendar year. Moreover, to the best of our knowledge it is the first time geostatistics is implemented as an effective tool in a cross-disciplinary solar irradiance analysis in a double role; that is, both for energy resources assessment at this detailed space-time level, and as a critical component for planning and decision-making in the deployment of state-wide solar projects.

In the following, we begin with a descriptive overview of our proposed framework in §2, alongside with informa-

tion about our case study in the state of California. Then, §3 provides a closer view to the collaborating methodologies in our framework prior to stepping through our analysis in detail. In §4 we discuss the results of this work, followed by our conclusions.

2. A Comprehensive Outlook

We approach the topic of selecting the number and installation locations for an MS network by introducing a 2-segment analysis framework. Our proposed framework is the composition of two complementary segments that combine cluster analysis schemes and geostatistics. The first segment is comprised by Optimal Cluster Selection (OCS), which is a clustering process and validation methodology to select an optimal number of clusters (NC) from an initial given set of GHI measurements. NC is the number of MS to be installed, by assigning a MS to each cluster area. Therefore, selecting the NC depends, for example, on the available resources to be invested for the creation of a MS network. Given the level of resources, a client can specify a feasible numeric window of MS. In response, OCS provides the analysis to determine an optimal NC, where optimality is deemed in terms of clustering validity assessment by adhering to statistical coherence features. In all, the OCS analysis improves MS placement by determining a structured station network scheme on the basis of (i) cluster analysis, and (ii) internal coherence criteria rooted in the cluster geometry.

OCS begins by performing cluster analysis of GHI temporal vectors that represent the solar irradiance temporal activity over the nodes of a gridded domain. The clustering algorithm employed in this study is the Adaptive Affinity Propagation (adAP), and yields a potential range of NC controlled by the algorithm's convergence criteria. From this range of NC we determine the optimal NC using a knee point detection scheme [26] on scientific criteria that pertain to cluster validity assessment [27, 28]. Namely, in the following we estimate the change of gradient (knee-point) of an evaluation graph of clustering quality metrics calculated through the measures of compactness and separation among the derived clusters.

Using the above scientific criteria, OCS produces possible MS configurations on the basis of the derived spatial segmentation into the optimal NC. For illustration, our work examines 3 such different location configurations; (A) the exact cluster centers as the obtained from the clustering process, (B) the clusters centers indicated by the median vector among each cluster's vector set, and (C) a random selection of location within the region of each cluster. Clearly, the concept of determining the alternative cluster locations in B and C is limited by the fixed area of a cluster, as well as the position of the exact center. We investigate additional characteristics of these configurations A, B, and C in the analysis Section §4. OCS makes configuration recommendations by examining the coherence of a cluster in terms of its intra-variance. In that sense, good quality clustering is expected to produce highly coherent cluster regions with low variance among the data members within a cluster. Consequently, we aim to explore the extent to which the derived clustering is able to capture clusters with low intra-variance and, thus, how representative the main cluster center is when compared to other candidate locations within the cluster.

After an optimal NC is determined and candidate MS configurations are suggested, the OCS methodology results are wired to the second analysis segment of our framework. The second segment uses geostatistical spatiotemporal analysis to sift through candidate MS configurations, and produces the optimal spatial MS distribution on the basis of geostatistical criteria. In particular, we partition the GHI data set into two subsets, namely a modeling and a validation subset. Given the abundance and the regular spatial distribution of the GHI observed data, our study assigns every other neighboring observation to be part of the first (modeling) subset, and places the remaining observations in the second (validation) subset.

In a first step, the modeling data subset is used as input values to predict GHI at the candidate MS locations. In a second step, we switch the role of the MS locations into GHI sampling locations as will be in real-life conditions. To this end, we consider the GHI predictions of the modeling first step (Step 1) as the input measurements for the second step (Step 2) of validation. Then, in Step 2, we use the MS locations input to predict GHI at the locations of the validation subset. We repeat these 2 steps for each one of the candidate MS configurations. In a third step

(Step 3), for each MS configuration we compare the accuracy of predicted GHI at the validation locations to the corresponding validation subset values. The validation results determine the optimal MS installation locations, according to the optimality criterion of prediction minimum error across space and time from the prediction tasks in Step 2. In the following, we also see that our study computes different error measures to facilitate selection according to the importance of a specific statistical measure in a given solar project.

For our study, we use a relatively large-scale solar irradiance dataset from SolarAnywhere [29], which consists of satellite-derived GHI values (in W/m^2 units) in a regular grid domain. Specifically, the GHI observations are taken from the Enhanced Resolution dataset provided by SolarAnywhere. These observations are spatially distributed across the entire state of California on a grid of approximately 0.01° resolution in both latitude and longitude at each temporal instance. Satellite imagery is periodically available at regular temporal intervals every 30' on the hour and half-hour for the entire year 2010 (in our study, times are reported in a 24-hour format and refer to local times in the time zone of California). Information about the cloud cover conditions and other parameters of the atmosphere and the earth's surface are obtained from geostationary satellite images. SolarAnywhere uses the SUNY semi-empirical satellite-to-irradiance model developed in [4] to generate global and direct irradiance values for locations within the continental US and Hawaii. Validation experiments at several climatically distinct locations have shown that typical errors of the derived global and direct irradiance estimates are close to 5% and 10%, respectively [4, 5].

To slightly reduce computational cost in our conceptual, large-scale regional paradigm, we perform our analysis by upscaling the original grid domain at a marginally sparser level. In particular, median filtering was applied to every non-overlapping 5×5 window of the grid to assign a window's median value to the center node of the window. As a result, each node in our study grid domain represents the median GHI of an approximately 25 km^2 square region, which is a reasonably dense scale in the context of our state-wide application. Our calculations exclude nodes that correspond to water areas, in addition to nighttime timestamps where irradiance is nonexistent. The previous considerations led us render a spatial domain of a total of 16625 grid nodes across the state of California for every temporal instance. Of the 16625 observations at each instance, we assigned 8313 to the modeling subset and remaining 8312 values to the validation subset, in the fashion detailed when the modeling and validation subsets were introduced.

In light of the nontrivial uncertainty in our input data set, as detailed in the previous, BME emerges as the most suitable geostatistical methodology for our intended spatiotemporal analysis. The geostatistical analysis relies on studying the characteristics of spatiotemporal homogeneity and correlation among the GHI values. In our study, a prediction task accounts for these characteristics among spatial locations at all temporal instances within 1 day to predict GHI at unsampled space-time locations throughout the day. The continuity and gradual succession of incoming solar energy patterns on a day-to-day basis means that the characteristics that define the behavior of the GHI field of values remain relatively unchanged in the scale of days or even a few weeks. For this reason, we choose to study GHI throughout the year 2010 by selecting representatively the first day of each month to perform a prediction task. In overview, our geostatistical prediction segment comprises of 12 such prediction tasks in 2010 to account satisfactorily for the seasonal change of incoming solar energy patterns. Table 1 displays the number of temporal instances considered by each prediction task in our study.

Overall, the geostatistical analysis segment involves performing a spatiotemporal prediction task at each time instance over all daylight instances in a day. The daily prediction tasks must be performed separately for each one of the modeling Step 1 and the verification Step 2. The complete analysis requires repetition of the steps for each selected day in the 12 months in 2010. Finally, the entire set of the aforementioned prediction tasks is repeated for each one of the proposed MS configurations A, B, and C to obtain the comparison measures for Step 3.

3. Methodologies In Our Study

3.1. Irradiance Data Input and Clear Sky Index

In our experiments, an $M \times N$ matrix of GHI daytime satellite-derived observations is used to describe the irradiance activity over the territory of CA. As described in Section 2, the number $M = 16625$ corresponds to the

Table 1: Number of consecutive daylight 30'-distanced temporal instances for each one of the 12 days in 2010 on which geostatistical prediction tasks are performed. The table also shows the local time (in 24-hour display) of the first instance considered on each day.

Date	# of Instances	Time of 1 st Instance
January 1	17	8.00
February 1	18	8.00
March 1	20	7.30
April 1	22	7.00
May 1	25	6.00
June 1	25	6.00
July 1	26	6.00
August 1	24	6.30
September 1	23	6.30
October 1	21	7.00
November 1	19	7.30
December 1	17	8.00

observation pixels of the upscaled gridded study domain and the $N = 8234$ stands for the consecutive daytime GHI variables every 30' for the year 2010. Namely, a row of the input matrix represents the temporal evolution of daylight GHI values during the period Jan 1 2010 to Dec 31 2010. Any days containing missing values have been utterly removed.

In solar engineering it is common to normalize the measured irradiance data over clear sky values, that is, the expected irradiance for clear sky conditions [30]. This preprocessing step calculates the so-called clear sky index attribute, which is used to eliminate the effects caused by diurnal and seasonal cycles. In this study, we use the clear sky model (CSM) developed in 2002 by Ineichen and Perez [31] which requires the Linke turbidity as the only atmospheric input. The clear sky index k_t for a time t at a given location is defined as the ratio:

$$k_t = \frac{I_t}{I_t^{CS}}, \quad (1)$$

where I is the satellite-derived GHI and I^{CS} is the calculated clear sky irradiance value. The k_t values range in $[0, 1]$, where 0 implies completely overcast sky and 1 stands for entirely clear sky conditions. For the cluster analysis, the original $M \times N$ observations matrix is transformed to a k_t -based content format.

3.2. Cluster Analysis Component

In this subsection we present the clustering framework for grouping the domain of interest into spatial coherent subdomains. We begin with a description of the clustering algorithm, followed by methodological descriptions of the clustering validation step and the determination of the optimal NC based on criteria related to clustering quality.

3.2.1. Adaptive Affinity Propagation

Affinity Propagation (AP) [32] is an extensively used clustering algorithm that appears in a number of data engineering applications; recently, AP has been also used in solar data clustering [11]. The distinguishable characteristic of AP against other unsupervised clustering algorithms is the fact that it is not based on random initial selection of k centers like other centroid-based algorithms. Instead, AP considers simultaneously all the data points as potential cluster centers (or *exemplars*). The concept of AP is iterative exchange of applicable messages between the exemplars until convergence is achieved towards a final clustering partition. The final exemplars are associated with maximization of the overall sum of similarities between the data points assigned to each cluster and their exemplars; that is

known as the maximization of the *net similarity*.

In particular, every node of the graph composed by all data points of the dataset, starts as a candidate exemplar. A similarity matrix S is required by AP as an input, which can be seen as a measure of similarity between two nodes of the dataset graph. Increased similarity between two points implies a strong possibility that one data point might be the exemplar to the other.

Two types of messages pass iteratively between the exemplars, namely, the *responsibility* $r(i, k)$ and the *availability* $a(i, k)$. The former represents the accumulated evidence about suitability of point k to serve as the exemplar of point i ; the latter indicates how appropriate it would be for point i to assign point k as its exemplar. The AP converges after a fixed number of iterations or when the message values fall below a threshold. As a result, each point is assigned to a cluster with point k as its exemplar, according to maximization of the following criterion:

$$k^* = \arg \max_k (a(i, k) + r(i, k)). \quad (2)$$

Although AP does not require a user-specified NC, the final clustering solution is influenced by a shared input parameter called the *preference*. *Preference* is defined as the self-similarity $s(k, k)$ of a point k , and corresponds to the appropriateness of point k to be chosen as an exemplar. The most common approach suggested by [32] is to set a common value of *preference* for all the points k , such as the median or the minimum of the similarity matrix S . However, the selection of the common *preference* value affects the final NC depending on how appropriate the *preference* is. Another parameter that could lead to a different clustering solution is the *damping factor* of AP. The *damping factor* is a parameter associated with regulation of AP convergence problems due to algorithmic oscillations when a stable solution cannot be reached. In this case a manually specified *damping factor* is used to update the *responsibility* and *availability* rules to alleviate the problem and improve the AP convergence performance.

Wang *et al.* [33] proposed an alternative version of AP, called Adaptive Affinity Propagation (adAP), which addresses the above limitations. The adAP performs (i) an adaptive search within the range of *preferences* (i.e., those that lead to the minimum and the maximum NC) in order to determine the optimal range on NC, and (ii) an adaptive process to automatically adjust the *damping factor*, as well as eliminate oscillations in case the damping method fails. The obtained range of number of clusters $[NC_{min}, NC_{max}]$ is much more restricted than the potentially available range of NC that theoretically varies between 1 and M clusters. Hence, the limited range can be seen as the spectrum of possible and stable solutions of an adaptive process focused on a specific dataset of attributes.

3.2.2. Clustering Feature

As mentioned in the previous subsection, the similarity matrix S plays an important role in the way of measuring the similarity between objects. Euclidean distance and Pearson correlation are the two most common measures, which are based on different conceptual approaches. The Euclidean distance between two nodes in AP can be defined by their real-value negative square Euclidean distance, and it is used to contrast the actual difference in values of attributes. On the contrary, the Pearson correlation coefficient is more appropriate to measure the similarity of the profile patterns of the objects, and is given by:

$$R_{i,j}|_{i \neq j} = \frac{cov(k_t(i), k_t(j))}{\sigma(k_t(i)) \sigma(k_t(j))}, \quad (3)$$

where $i, j = \{1, \dots, M\}$ and $k_t(\cdot)$ are rows of the $M \times N$ normalized GHI matrix.

Indeed, we consider using the Pearson correlation as more appropriate for capturing the correlation of the temporal expression profiles of solar irradiance activity, in terms of capturing the prevailing variability of GHI. In other words, solar engineering is interested in grouping spatial regions according to similar variability within a cluster; thus, this is revealed by the shape correlation of the solar patterns per observation pixel, in the manner of clustering large-scale gene expression data [34, 35].

3.2.3. Cluster Validity Framework

From a solar engineering viewpoint, the derived range of number of clusters $[NC_{min}, NC_{max}]$ can be considered as an affordable and feasible range on NC for an operational provider company to invest in a given service area. In the context of a typical unsupervised cluster analysis, it is a valuable finding to define a recommended NC within the available range of clustering solutions. To this end, a number of internal cluster validity indices (VI) have been proposed to estimate the quality of each cluster partition, in terms of compactness and separation of the obtained clusters [28]. The most well-cited VI have been extensively used in solar irradiance data clustering [9, 10, 11]. Research in these studies shows that the Calinski-Harabasz (CH) [36] VI performs smoothly and provides a monotonically decreasing evaluation graph as the NC increases. Based on this finding, it can be argued that using CH to determine the recommended NC for an application is preferable to the Silhouette [37] VI adopted in [33]. The CH index is a normalized ratio-type expression of the "between cluster scatter" to the "within cluster scatter", which estimate inter-cluster separation and intra-cluster cohesion, respectively. According to this approach, larger CH values are associated with increased clustering quality.

By calculating the CH values for each clustering case in the range $[NC_{min}, NC_{max}]$, we obtain the evolution of the validation metric with respect to the number of clusters. A knee-point detection method can be used to determine the change of gradient of the graph, which is considered as the graph point with the highest critical role. The use of a graphical method to detect a knee-point is called *L-method*, and was introduced in [26]. It was first adopted by Zagouras *et al.* in several solar engineering studies regarding the determination of optimal NC [9, 10, 11]. The *L-method* applies on the evaluation graph to fit a pair of straight lines on consecutive points of the graph curve that separate the curve into two segments. An iterative process examines all possible pairs of fitted lines in terms of calculating the total root-mean-square error (*RMSE*) of the overall fit. The pair of lines with the minimum total *RMSE* is the best fit, and the point of intersection of the two lines indicates the knee-point. More technical details regarding the *L-method* can be found in [26, 10]. In essence, we diverge from the clustering evaluation approach proposed by Wang *et al.* [33], in that we address determination of optimal NC of solar irradiance data in a more focused manner.

3.3. Geostatistical Component: Analysis with the Bayesian Maximum Entropy

As detailed in §2, geostatistical analysis is used in two sequential steps; namely, for modeling GHI at the proposed MS locations, and for validation of the MS-based prediction in the state-wide grid of the validation data subset. To apply geostatistics, well-known methodologies exist in literature. Most approaches revolve around the family of kriging methods; see, e.g., [38], [39], [40], and [41]. The Bayesian maximum entropy method (BME; [42], [43]) extends common geostatistical approaches by operating in a knowledge synthesis framework that is presented in the following. In brief, the core advantages of BME over mainstream techniques are two-fold: (i) BME does not require common restrictive assumptions made by other geostatistical approaches, such as the requirements for Gaussianity and linearity of space-time data, and (ii) BME is unique among other methods in allowing rigorous incorporation of data uncertainties into the stochastic prediction process.

The BME knowledge synthesis (KS) framework was proposed by Christakos in [43]. According to the KS framework, our knowledge about space-time solar irradiance can be classified into the following two major categories of knowledge bases:

- The General knowledge G-KB, which reflects our level of knowledge regarding main characteristics of solar irradiance, such as statistical measures about its mean and covariance, and scientific facts that apply in the analysis context. G-KB is derived from epistemic principles, such as physical laws, conceptual models, and statistical moments.
- The Site-specific knowledge S-KB, which involves the solar irradiance observations and data measurements at specific spatial points and time instances. S-KB can exhibit various confidence levels of data accuracy; for example, single values characterize hard, exact data for which we are certain or carry trivial uncertainty, whereas interval and probabilistic forms are used to express soft data with measurable, nontrivial uncertainty.

Solar irradiance as an attribute has a stochastic nature, justified by inherent uncertainties, its spatiotemporal heterogeneity, and the imperfect nature of our knowledge about it. According to the BME approach, solar irradiance on the Earth's surface can be thus characterized effectively by representing it as a spatiotemporal random field (S/TRF) $X_p = X_{s,t}$ at each location $p = (s, t)$ in the unified spatial and temporal continuum with spatial coordinates $s = (s_1, s_2)$ and temporal coordinate t [44]. Specification of the X_p values at all points of the continuum determines a realization of the S/TRF. Randomness manifests itself as an ensemble of possible realizations of the X_p distribution. BME incorporates the two knowledge bases G-KB and S-KB to blend all available information about solar irradiance according to the following pair of equations:

$$\begin{aligned} \int d\chi (\mathbf{g} - \bar{\mathbf{g}}) e^{\mu^T \mathbf{g}} &= 0 \\ \int d\chi \xi_s \mu^T \mathbf{g} - A f_K(\chi) &= 0 \end{aligned} \quad (4)$$

where χ are the realization values of the S/TRF, $\mathbf{g}$ is a vector of g_α -functions ($\alpha = 1, 2, \dots$) that represents stochastically the available G-KB (the bar denotes statistical expectation). In Eqs. 4, μ is a vector of μ_α -coefficients that depends on the space-time coordinates and is associated with $\mathbf{g}$. Specifically, the μ_α express the relative significance of each g_α -function in the composite solution. The maximum entropy principle is employed to enable incorporation of as many relevant G-KB sources available in an initial prior stage expressed by the first of Eqs. 4. This stage leads to the prior probability density function (PDF) $e^{\mu^T \mathbf{g}}$ that is constrained by the G-KB, as shown in the first equation of Eqs. 4. The second equation of Eqs. 4 represents succinctly the second stage in BME analysis, known as the posterior or integration stage, where ξ_s represents the S-KB available, and A is a normalization parameter. The integration stage uses operational Bayesian updating theory [45] to conditionalize the site-specific S-KB on the prior G-KB. By doing so, it effectively blends all available information about solar irradiance to produce posterior PDFs, designated as f_K , across all space and time locations we seek results at. The $\mathbf{g}$ and ξ are the inputs in Eqs. 4, whereas the unknowns are the μ coefficients and f_K across space-time.

BME is an extension of classical geostatistical methodologies, in the sense that these methods can be described as limiting cases of the BME methodology. For example, when the available G-KB is the X_p mean and covariance and the S-KB comprises only hard data, then BME reduces to the ordinary kriging predictor. Additional information about BME applications and related software can be found online at spacetimeworks.com, whereas for a BME presentation in exhaustive detail see [43].

4. Analysis Of Irradiance Data

4.1. Clustering Map

We computed the similarity matrix S of the normalized $M \times N$ GHI matrix, based on the Pearson correlation coefficient; the goal is to create inputs comprised of correlations between the solar irradiance time series of each observation pixel and all other pixels. Next, the adAP was used to perform clustering into spatial groups of similar solar irradiance profiles. We used the MATLAB code for adAP from [46]. The clustering process suggested a range of meaningful NC that varies between $[NC_{min}, NC_{max}] = [59, 545]$ clusters. This resulting range emerged from an adaptive, iterative clustering process that shrunk the initial number of M potential exemplars to 545, in addition to computing the minimum reasonable NC to be 59. On the basis of adAP, any NC within this range can be considered as a valid recommendation according the adAP convergence criteria.

Next, we took an additional step to propose an optimal NC by means of a cluster validity estimation. Specifically, the *L-Method* was applied to the evaluation graph of the CH index, and determined that no significant change of the validity index occurs for NC=189. Figure 1 illustrates the clustering map of the 189 coherent spatial clusters produced by this process. The shape and boundaries of each cluster are precisely depicted with regard to the homogeneity of correlation between the data members in each region.

Figure 1: Map representation of the clustering structure of California from the Adaptive Affinity Propagation analysis. The number of coherent spatial clusters of GHI observations is determined to be $NC=189$ throughout the area of the state, as validated by the CH validity index. The colorbar indicates assigned cluster labels to each spatial group of pixels.

4.2. MS configurations

Upon computing the optimal NC , from a clustering perspective all spatial locations inside a cluster are appropriate candidate ground sites to install the MS that corresponds to this cluster. For illustration, Figure 2 depicts an example of three possible configurations of MS locations in a cluster. A brief description of these MS configurations follows:

Set A: The MS are placed at the computed cluster centers as obtained by the clustering algorithm.

Set B: The MS are placed by determining the vector median of each cluster's vector set. Assume that $\mathbf{X}_C = \{\vec{x}_r\}$ is the set of n_C vectors that forms the group of vectors within the same cluster C . Then, the aggregate distance of each vector $\vec{x}_i$ from the other vectors in the set is defined as:

$$d_r = \sum_{i=1}^{n_C} (\|\vec{x}_i - \vec{x}_r\|)_p, \quad (5)$$

where $\|\cdot\|_p$ stands for the vector norm; in our analysis, we consider the vector norm to be the Euclidean 2-norm. For each cluster vector set, we calculate all the d_r distances. The first smallest ranked distance is used to define the vector median (VM) [47]. In other words, the VM is the representative among vectors in a set on the basis of the minimum aggregate distance from all vectors in this set.

Set C: The MS are alternatively placed at random points within a specified range R from the initially computed cluster centers in Set A. Although the selected location within R need not be random, in our study we choose to apply a random selection for each cluster. Our choice serves to illustrate the possibility that external factors, such as terrain and location characteristics, might render unfeasible installing the MS at the exact cluster centers. This choice also plays the role of a control-case that involves random positioning, as opposed to the highly organized positioning of Sets A and B. For our implementation, we clearly need to select a random location that places a MS within the boundaries of the cluster it represents. With respect to the minimum distance between the furthest pixel of any cluster and its cluster center, an alternative cluster center is randomly selected by choosing one pixel for each cluster that is within a distance range of $25km \leq R \leq 50km$ from each initial cluster center (Set A). Note that in different test cases, additional analysis could possibly suggest different optimal value for

R according to case-specific criteria; hence, with respect to MS placement, our framework provides significant level of flexibility for real-world applications.

Figure 2: Example of three different MS configurations within a cluster that are used in our study. The figure shows MS positioned at the cluster main center, the VM center, and a random center. The VM center is usually located close to the main center as they both deal with the minimization of distance-type associations to the other cluster members. In the case of the random centers, the geometrical dimensions of the derived clusters are observed; specifically, the random center is selected to be at least 25km away from the main center, and at maximum distances that remain within the boundaries of the cluster.

In our present proof-of-concept study, we illustrate our framework by proposing the above 3 MS installation configurations A, B, and C for the state of California. Observe that even in the case of non-random placement of MS, it might be physically intangible to build stations at proposed locations due to the presence of natural obstacles, geographical constraints, terrain anomalies, etc. To this end, our analytical scheme recommends replacing locations with the immediately lower-ordered spatial configuration locations that are physically tangible.

4.3. Geostatistical GHI Prediction

According to the preceding clustering segment, geometrical and statistical cluster criteria suggested installing 189 MS in 3 different configurations A, B, and C, as detailed in the previous. In the following, we develop an approach to complement the cluster analysis by predicting GHI from uncertain data beyond the MS locations across the state of California at the 30' time resolution, and further by investigating whether any of the proposed MS configurations can be recommended as optimal according to geostatistical criteria.

In each one of the BME geostatistical modeling and verification Steps 1 and 2, respectively, that were presented in §2, we have soft data as input. In particular, Step 1 takes as input the modeling subset of satellite data; we consider the entirety of those as soft intervals where the true value might lie anywhere in the interval defined by the uncertainty levels of satellite irradiance observations, as detailed in §2. In the prediction tasks of Step 2, the input is again entirely soft data in the form of Gaussian distributions; their mean is the BME mean computed in Step 1, and their variance is the corresponding prediction variance from Step 1. In each one of the Steps 1 and 2, the corresponding input data are initially examined for surface trends. These trends express the GHI S/TRF behavior at larger scales and can overshadow the S/TRF structural characteristics at the present scale of study. For this reason, we compute estimates of the trends by swiping an exponential kernel smoothing window across California at each spatiotemporal instance, and then remove the trends prior to BME prediction. The trends are subsequently reinstated after prediction to obtain nonresidual GHI values.

For illustration, Figure 3 shows a sample of the estimated trends at two consecutive instances $t = 3$ (07.30), and $t = 4$ (08.00) on June 1 2010 for both Steps 1 and 2 of the configuration A analysis. In the figure, subplots Figure 3a

and Figure 3b refer to Step 1, so the trend estimates are shown at the Step 1 prediction locations which are the proposed MS spatial locations. Correspondingly, subplots Figure 3c and Figure 3d refer to Step 2, so the trend estimates are shown at the Step 2 prediction locations which are the 8312 validation points across California. The GHI color scales in the plots of each Step analysis are common for all plots, and account for the GHI minima and maxima across the entire study period in 2010 to facilitate comparison.

Figure 3: Estimated trends at prediction locations based on the analysis of MS configuration A at selected time instances on June 1, 2010. All plots in W/m^2 units; coordinates are measured in decimal degrees (DD). (a) and (c) show estimated trends in the modeling Step 1 at time instance $t = 4$ (07.30); (b) and (d) show estimated trends in the validation Step 2 at time instance $t = 5$ (08.00).

1

The residual GHI is analyzed for internal structure patterns at a series of spatial and temporal distances, and a spatiotemporal distance-based empirical correlation is computed. A suitable function is then fitted to the empirical curve to yield the covariance model that describes the GHI S/TRF characteristics on a specific day. Covariance models are comprised of simple or nested functions that must comply with the permissibility criterion of positive definiteness [38]; commonly used functional forms that abide to this rule are the exponential, Gaussian and spherical form. Spatiotemporal covariances typically consist of separable or nonseparable spatial and temporal components, and can be constructed from such simple permissible functions, or can be more elaborate permissible algebraic constructs derived from general principles, physical-based models, etc., as shown by Kolovos *et al.* in [48]. Eventually, the fitted covariance is used as G-KB in BME prediction, alongside with the zero-mean GHI residuals.

A limited preliminary analysis [49] supported our earlier claim in §2 that the structural characteristics of the GHI S/RTF can be considered relatively unchanged in the temporal scale of a few weeks. To this end, we used the same covariance function for all prediction tasks throughout the time instances of a single day. Furthermore, we computed anew the covariance function for each different day across the 12-day period in 2010. The empirical covariance is estimated from the input data, so in Step 1 we use the same input data set of the 8313 modeling subset data for all MS configurations A, B and C. In Step 2, the input data are situated at the MS configuration locations. The slightly different locations of the input data in each configuration led to empirical covariances that were often different across the A, B and C configurations. As a result, in the validation Step 2 the covariance models used for prediction vary not only per month but also per configuration, too. A discussion about the covariance functions used in the present geo-statistical study is provided in further detail in Appendix A. For illustration, Figure 4 shows plots of the covariance functions used for June 2010 in Step 1, and a selected configuration (C) in the same month in Step 2.

Figure 4: Empirical and modeled covariance plots for the day of June 1, 2010. The vertical covariance axes are in $[W/m^2]^2$ units; the horizontal plane shows the spatial lags (in decimal degrees) vs. the temporal lags (in 30-minute increments). (a) shows the covariances for the modeling Step 1; (b) shows the covariances for configuration C of the the validation Step 2.

In Step 1, BME prediction produces the posterior GHI PDFs at the proposed MS locations for each one of the examined configurations A, B, and C. In Step 2, we take the BME means of these predicted PDFs and the standard prediction errors to compose the input data for the validation task. Observe that as in Step 1, so in Step 2 the entire input data set consists of soft data distributions that represent realistically the expected data and their uncertainty to be measured at the proposed MS locations, regardless of the configuration scheme.

The prediction tasks in Step 2 perform an extensive validation process, in which BME prediction at the 8312 validation points across California and at each temporal instance yields a measurable size of GHI values to be compared up against the original subset of validation data. In paper, we can only provide a sample of the BME prediction volume of output through the following illustrations. Several animations showing the entire BME prediction output for all configurations on each one of the 12 days examined in 2010 can be found in the Electronic Supplementary Content. As an example, Figure 5 shows the BME mean of the posterior PDFs at selected instances for all configurations A, B, and C. The plots display GHI values after the mean trends have been restored, whereas all BME mean plots are presented at the same time instances and in the same color scale to facilitate comparison. By inspecting the output at a specific instance, results indicate that very similar GHI patterns are obtained by using any of the A, B, or C MS configurations. The prediction differences define the behavior of the validation errors, as shown in the following section.

With respect to the prediction error, Figure 6 exhibits the BME prediction standard error for selected instances in the configuration A analysis. The plots suggest that standard error is expectedly minimal close to the data locations, whereas it can increase up to about 40–50 W/m^2 at locations that are farthest away from the data. Unlike mainstream geostatistical methodologies that report prediction error values based only on the prediction task, the BME prediction error conveys a more accurate picture about prediction uncertainty because it also incorporates the initial error embedded in the input soft data.

4.4. Performance Notes

In terms of performance, the computational burden for the AP algorithm primarily depends on the selected number of clusters to operate on; for BME, the controlling factors are the data set size defined by the $M \times N$ product, and the number of prediction nodes. We note that performance of the AP algorithm is independent of the parameters M and N , and that applying AP yielded results in the temporal order of minutes. However, application of BME in the scale of our study demanded cumulatively about a day of execution time to produce results for each of the Set A, B, and C cases on a 6-core Intel® i7 3.4 GHz desktop computer with 32 GB of RAM. Although BME poses no inherent limitations on the size of M and N parameters, the computational burden can be expected to scale with increasing M and N values in particular due to the geostatistical search for nearest neighbors among increasingly larger data populations. Clearly, the BME computational burden also increases when prediction is sought at additional space-time locations.

4.5. Comparison of Proposed MS Configurations

To inspect the proposed MS configurations A, B, and C in the final Step 3, we use a series of error measures to compare the BME predicted GHI $\hat{\chi}$ at the validation locations in Step 2 to the actual values χ of the validation data subset at the corresponding locations. For a sample of N_d data, prediction accuracy is measured by computing the following error statistics:

- Mean Absolute Error (*MAE*, in W/m^2 units), defined as:

$$MAE = \frac{1}{N_d} \sum_{i=1}^{N_d} |\chi_i - \hat{\chi}_i| \quad (6)$$

- Mean Bias Error (*MBE*, in W/m^2 units), defined as:

$$MBE = \frac{1}{N_d} \sum_{i=1}^{N_d} \chi_i - \hat{\chi}_i \quad (7)$$

- Mean Relative Error (*MRE*, percentage), defined as:

$$MRE = \frac{1}{N_d} \sum_{i=1}^{N_d} \frac{\chi_i - \hat{\chi}_i}{\chi_i} \quad (8)$$

- Root Mean Square Error (*RMS E*, in W/m^2 units), defined as:

$$RMS E = \sqrt{\frac{1}{N_d} \sum_{i=1}^{N_d} (\chi_i - \hat{\chi}_i)^2} \quad (9)$$

- Coefficient of determination (R^2 , dimensionless); given $\bar{\chi} = \frac{1}{N_d} \sum_{i=1}^{N_d} \chi_i$, R^2 is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{N_d} (\chi_i - \hat{\chi}_i)^2}{\sum_{i=1}^{N_d} (\chi_i - \bar{\chi})^2} \quad (10)$$

By sweeping across all validation spatial locations, we compute the above measures in 3 different ways:

1. Global total error across all temporal instances. This metric provides the error distribution throughout an annual cycle in California as a summary about the performance of each proposed MS configuration.
2. Solar noon error throughout the year. This metric aggregates the errors over the entire year and compares the error distributions at the time instance $t = 12.00$, which we consider to be the average solar noon across the state of California. This perspective can indicate whether MS configuration performance might be dependent on the time of the day.
3. Seasonal error across trimesters in the year. This metric provides the error on a seasonal basis for the months in winter (January, February and December), spring (March–May), summer (June–August), and fall (September–November), and can suggest whether a proposed configuration might have an advantage over the rest during a specific season of the year.

The results of the above error computations are presented in Table 2. A first comparison indicates interestingly that there is no clear winner among the proposed MS configurations A, B and C. Regardless of the criterion of choice among the statistical measures in Table 2, all types of placements for the MS appear to perform about equally well with only small relative differences among error values in the corresponding measure. Also, it appears that no single error computation among the global, solar noon and seasonal types could act as a preferred assessment filter to select an MS configuration as a clear winner over the rest. However, there is some consistency across the different computations in that configurations A and B both exhibit slightly better performance than C.

As yet a different error measure, Figure 7 shows the distribution of the validation Step 2 relative error from all spatiotemporal locations for all 3 proposed MS configurations A, B, and C. The sample of measurements used to produce each one of those distributions is well above 2.1 million spatiotemporal locations, as shown in Table 3. The figure includes the counts of all relative errors $\leq |100\%$. A small percentage of less than 5% of locations across each MS configuration A, B, and C presented relative errors $> |100\%$. Table 3 has more detail about the percentage of relative errors $\leq |10\%$ and $> |100\%$ per MS configuration.

In brief, about $2/3$ of our predictions in validation Step 2 have relative errors $\leq |10\%$. This can be considered a satisfactory result, given (a) the amount of spatiotemporal locations in the validation grid that provide a reliably large sample for our statistical computations, (b) the relatively small amount of MS locations that play the role of input data locations for the prediction tasks in Step 2, and (c) the nontrivial uncertainty in the entire data set of GHI measurements. In these conditions of increased uncertainty, our combined clustering and geostatistical analysis indicates a marginally better performance of MS configurations A and B against random placement configuration C. According to the criterion of having the majority of relative errors nearest to zero, Figure 7 and Table 3 suggest there is a borderline advantage in using the MS configuration B against A.

5. Conclusions

Our work proposes an adjustable framework to select locations for solar measuring stations on the basis of spatial feature coherence, geometrical and geostatistical criteria of optimality. It combines the strength of established, powerful methodologies in cluster analysis and geostatistics to illustrate a cross-disciplinary research paradigm for the benefit of renewable energy monitoring and generation.

We apply this framework in a relatively large-scale study of MS placement in California. To this goal, we utilize a high-resolution solar irradiance satellite dataset that covers the area of interest across an entire calendar year. These satellite-derived GHI data are initially the input to a cluster analysis method. Cluster validity criteria lead to an optimal number of clusters that constitute a clustering map of similar, in shape, solar irradiance profiles during an annual testing period. The Adaptive Affinity Propagation methodology is then employed to conduct clustering and to address clustering algorithm limitations in parameter selection. Upon computing the optimal spatial segmentation of coherent groups, we examine 3 different scenarios of representative MS configurations. The clustering results are then subjected to geostatistical analysis to further investigate optimal locations in the clusters for the MS installation on the basis of geostatistical criteria. This analysis consists of modeling assumed GHI measurements at the configuration

Table 2: Comparative presentation of errors from geostatistical validation analysis (Step 2) of GHI in California for the 3 proposed MS configurations A, B, and C. The table displays the Mean Absolute Error (MAE), Mean Bias Error (MBE), Mean Relative Error (MRE), Root Mean Square Error ($RMS E$), and Coefficient of determination (R^2) for global, solar noon, and seasonal error computations.

	A	B	C
Global			
MAE	49.1104	48.7825	50.6951
MRE	-11.8571	-11.8839	-12.2981
MBE	2.4315	1.7159	3.2798
$RMS E$	78.1945	78.361	79.1633
R^2	0.92489	0.92457	0.92302
Solar Noon			
MAE	63.0742	62.5414	65.1517
MRE	-4.4475	-4.4108	-4.6251
MBE	14.5083	13.743	15.1532
$RMS E$	100.1871	100.2872	101.3306
R^2	0.87007	0.86981	0.86708
Seasonal			
December–February			
MAE	47.3206	47.1258	48.2812
MRE	-17.9833	-17.5768	-20.2124
MBE	1.0414	1.2065	-1.179
$RMS E$	65.552	65.7187	66.0965
R^2	0.86405	0.86336	0.86178
March–May			
MAE	59.7437	59.1394	62.1031
MRE	-12.4534	-12.7014	-12.1041
MBE	1.3141	-0.28176	3.8079
$RMS E$	91.0393	91.0734	92.7486
R^2	0.89275	0.89267	0.88869
June–August			
MAE	44.7449	44.7812	46.3756
MRE	-6.5341	-6.6568	-6.2522
MBE	3.376	2.8597	5.9421
$RMS E$	75.2597	75.9025	75.4495
R^2	0.93289	0.93174	0.93255
September–November			
MAE	44.3311	43.7556	45.5473
MRE	-12.4278	-12.4646	-13.086
MBE	3.6775	2.9369	3.2711
$RMS E$	76.3347	76.1131	77.5905
R^2	0.90945	0.90998	0.90645

Table 3: Percentages in the distribution of relative errors (RE) computed in Step 2 where the relative error is $\leq |10|\%$ or $> |100|\%$ for the 3 proposed MS configurations A, B, and C.

	A	B	C
Validation observations used	2119546	2119549	2119560
% of validation RE $\leq 10 \%$	59.5533	59.8786	58.9548
% of validation RE $> 100 \%$	4.4495	4.4295	4.4578

locations, then predicting GHI at an extensive set of validation locations, and finally analyzing a series of statistical errors from the validation step.

Geostatistical analysis of the GHI data that consists entirely of uncertain measurements depends profoundly on the unique features of BME among other geostatistical methodologies. Not only does BME provide insight about the GHI spatiotemporal random field stochastic structure, but it also accounts for the known, nontrivial GHI measurements uncertainty in a rigorous manner that integrates the measured uncertainty in predicted values. By employing BME in our framework we achieved a two-fold goal which is to (a) extend existing work of geostatistical GHI modeling by applying prediction in a larger-scale spatial domain at subhourly intervals, and (b) provide an assessment tool to complement the cluster analysis towards selecting installation locations for GHI measuring stations.

Our study concludes that overall all 3 proposed MS location configurations score statistically almost equally well under the geostatistical analysis criteria. This finding suggests an element of robustness in the cluster analysis, since all of the clustering criteria-proposed configurations fare almost as well under the additional geostatistical selection filter. The nearly-even performance of all configurations across different statistical measures also suggests strongly coherent intra-cluster variance in each cluster, which stands in support of the approach we have developed to define the clusters and propose MS installation locations.

Subsequently, more than offering a choice among proposed locations in an MS construction project, our study enables the liberty to select the locations from any configuration that would perform best according to possibly additional case-specific criteria. Such criteria might be related to, e.g., geographic and terrain conditions, or economic considerations that might stem from a specific MS configuration choice. Alternatively, a selection can be made according to the gravity of any of the specific statistical measures analyzed in §4.5. For example, we illustrated that under the minimum relative error criterion the MS configuration A is marginally preferable to configurations B and C. In any scenario, the validation error analysis can serve as a quantitative tool to assess optimality in selecting MS network locations. Our proposal goes beyond the narrow boundaries of a focused application, and aims to help specialists across the spectrum of the renewable energy industry for planning, decision-making, and policy implementations.

Acknowledgments

Dr. Zagouras was supported by a California Solar Initiative Grant from the California Public Utilities Commission, and Prof. Coimbra is funded by the National Science Foundation, the California Public Utilities Commission, the California Energy Commission and by the U.S. Department of Energy.

Appendix A. Covariance Modeling

In geostatistical analysis, spatiotemporal covariances provide the BME predictor with structural information about spatial homogeneity and temporal stationarity of the GHI S/TRF X_p . In our case, we study zero-mean GHI residuals for which we also obtain second order statistical moments empirically from the data. In the absence of additional information about the GHI random field, we consider a uniform variance for X_p . According to these considerations, X_p is second-order homogeneous in space and stationary in time, in which case covariance depends only on the distance $\mathbf{h}$ between any pair of spatiotemporal locations.

Specifically, between any two points $\mathbf{p}_1$ and $\mathbf{p}_2$, the space-time distance between them is $\mathbf{h} = \Delta\mathbf{p} = \mathbf{p}_2 - \mathbf{p}_1 = (s_2 - s_1, t_2 - t_1) = (\Delta s, \Delta t) = (r, \tau)$. Selection of the spatiotemporal distance metric is critical for the correct computation of distances and covariances. Different measures might be appropriate for different problems, as shown in [50]. In our study, we use the common Euclidean distance metric to define spatiotemporal distances across California and the temporal instances within a day.

Our study uses ordinary covariance functions ([38]) which comprise one or more positive definite functional forms. Frequently used functional forms are

- the exponential covariance: $C_{Exp}(c, \alpha) = c \exp(-\mathbf{h}/\alpha)$, where $0 < \alpha$
- the Gaussian covariance: $C_{Gau}(c, \alpha) = c \exp(-\mathbf{h}^2/\alpha^2)$, where $0 < \alpha$
- the spherical covariance: $C_{Sph}(c, \alpha) = c [1 - \frac{3|\mathbf{h}|}{2\alpha} + \frac{|\mathbf{h}|^3}{2\alpha^3}]$, if $|\mathbf{h}| < \alpha$, and $= 0$, if $\alpha \leq |\mathbf{h}|$

where parameter c denotes the covariance sill that equals the S/TRF variance, and parameter α is the positive covariance range that indicates the spatiotemporal distance at which correlation extends within the S/TRF. Moreover, we use nonseparable covariance functions where the spatial and temporal components cannot be expressed in separate terms. One or more such products of spatial and temporal simple functional forms can be used to express a covariance function model that has been fitted to the empirical covariance. For the present study, we applied a visual fit, by following the comments in [51].

According to the previous, the general expression for covariances in our study is the sum of one or more nested products in the format: $C = \sum_i c_i C_{si} C_{ti}$, $i = 1, \dots, N_{nest}$, where typically $N_{nest} \leq 3$, and one sill parameter applies per term. For example, the covariance model plots in Figure 4 depict the following models:

Figure 4a: $C_{Step1, June1} = 4000 C_{r, Exp}(0.17) C_{\tau, Sph}(4) + 641 C_{r, Exp}(0.9) C_{\tau, Sph}(17)$

Figure 4b: $C_{Step2, ConfC, June1} = 3400 C_{r, Sph}(0.5) C_{\tau, Exp}(5) + 24 C_{r, Sph}(20) C_{\tau, Exp}(5)$

In the above, both models have 2 nested components each. By inspecting the sill values, in both models the first component contributes significantly to the total covariance. In the first model, spatial correlation extends in an exponential fashion to about 0.17 decimal degrees (DD) according to the range of the first spatial term, whereas the second spatial term has a much weaker effect of extending the correlation range to 0.9 DD also exponentially. Similarly, temporal correlation appears to extend primarily to 4 instances, that is 2 hours, around the current temporal instance, whereas the second temporal term extends temporal correlation more weakly to 17 instances which cover about the length of the entire daytime. In both cases, we found temporal correlation to be best illustrated by a spherical model form. The second model can be qualitatively described in a similar manner.

References

- [1] A. Ohmura, E. G. Dutton, B. Forgan, C. Fröhlich, H. Gilgen, H. Hegner, A. Heimo, G. König-Langlo, B. mcarthur, G. Müller, R. Philipona, R. Pinker, C. H. Whitlock, K. Dehne, M. Wild, Baseline surface radiation network (BSRN/WCRP): New precision radiometry for climate research, *Bull. Amer. Met. Soc.* 79 (1998) 2115–2136.
- [2] K. Scharmer, J. Greif, The European solar radiation atlas: Database and exploitation software, The European Solar Radiation Atlas, Les Presses de l'École des Mines, 2000.
- [3] M. R. Risch, D. M. Kenski, D. A. Gay, A great lakes atmospheric mercury monitoring network: Evaluation and design, *Atmospheric Environment* 85 (2014) 109–122.
- [4] R. Perez, P. Ineichen, K. Moore, M. Kmiecik, C. Chain, R. George, F. Vignola, A new operational model for satellite-derived irradiances: description and validation, *Solar Energy* 73 (5) (2002) 307–317.
- [5] R. Perez, P. Ineichen, M. Kmiecik, K. Moore, D. Renne, R. George, Producing satellite-derived irradiances in complex arid terrain, *Solar Energy* 77 (4) (2004) 367 – 371.
- [6] M. Román, C. Schaaf, C. Woodcock, A. Strahler, X. Yang, R. Braswell, P. Curtis, K. Davis, D. Dragoni, M. Goulden, L. Gu, D. Hollinger, T. Kolb, T. Meyers, J. Munger, J. Privette, A. Richardson, T. Wilson, S. Wofsy, The MODIS (Collection V005) BRDF/albedo product: Assessment of spatial representativeness over forested landscapes, *Remote Sensing of Environment* 113 (11) (2009) 2476–2498.
- [7] M. Hakuba, D. Folini, A. Sanchez-Lorenzo, M. Wild, Spatial representativeness of ground-based solar radiation measurements, *Journal of Geophysical Research: Atmospheres* 118 (15) (2013) 8585–8597.
- [8] E. G. Dutton, D. W. Nelson, R. S. Stone, D. Longenecker, G. Carbaugh, J. M. Harris, J. Wendell, Decadal variations in surface solar irradiance as observed in a globally remote network, *Journal of Geophysical Research: Atmospheres* 111 (D19).
- [9] A. Zagouras, A. Kazantzidis, E. Nikitidou, A. Argiriou, Determination of measuring sites for solar irradiance, based on cluster analysis of satellite-derived cloud estimations, *Solar Energy* 97 (2013) 1–11.
- [10] A. Zagouras, R. H. Inman, C. F. M. Coimbra, On the determination of coherent solar microclimates for utility planning and operations, *Solar Energy* 102 (2014) 173–188.
- [11] A. Zagouras, H. T. Pedro, C. F. M. Coimbra, Clustering the solar resource for grid management in island mode, *Solar Energy* 110 (2014) 507–518.

- [12] O. Prakash, B. Datta, Sequential optimal monitoring network design and iterative spatial estimation of pollutant concentration for identification of unknown groundwater pollution source locations, *Environmental Monitoring and Assessment* 185 (2013) 5611–5626.
- [13] F. Antonanzas-Torres, A. Sanz-Garcia, F. M. de Pisón, O. P. nán Lamigueiro, Evaluation and improvement of empirical models of global solar irradiation: Case study northern Spain, *Renewable Energy* 60 (2013) 604–614. doi:http://dx.doi.org/10.1016/j.renene.2013.06.008.
- [14] Z. Dong, D. Yang, T. Reindl, W. M. Walsh, Short-term solar irradiance forecasting using exponential smoothing state space model, *Energy* 55 (2013) 1104–1113. doi:http://dx.doi.org/10.1016/j.energy.2013.04.027.
- [15] G. Christakos, *Random Field Models In Earth Sciences*, Academic Press, San Diego, CA, 1992.
- [16] V. D’Agostino, A. Zelenka, Supplementing solar radiation network data by co-kriging with satellite images, *International J. of Climatology* 12 (1992) 749–761.
- [17] S. Rehman, S. G. Ghor, Spatial estimation of global solar radiation using geostatistics, *Renewable Energy* 21 (3–4) (2000) 583 – 605. doi:http://dx.doi.org/10.1016/S0960-1481(00)00078-1.
- [18] C. Ertekin, F. Evrendilek, Spatio-temporal modeling of global solar radiation dynamics as a function of sunshine duration for Turkey, *Agricultural and Forest Meteorology* 145 (2007) 36–47.
- [19] D. Yang, C. Gu, Z. Dong, P. Jirutitijaroen, N. Chen, W. M. Walsh, Solar irradiance forecasting using spatial-temporal covariance structures and time-forward kriging, *Renewable Energy* 60 (2013) 235–245. doi:http://dx.doi.org/10.1016/j.renene.2013.05.030.
- [20] G. Christakos, A. Kolovos, M. L. Serre, F. Vukovich, Total ozone mapping by integrating databases from remote sensing instruments and empirical models, *IEEE Trans. on Geoscience and Remote Sensing* 42 (5) (2004) 991–1008.
- [21] P. Bogaert, G. Christakos, M. Jerrett, H.-L. Yu, Spatiotemporal modelling of ozone distribution in the state of California, *Atmospheric Environment* 43 (2009) 2471–2480.
- [22] A. Kolovos, A. Skupin, M. Jerrett, G. Christakos, Multi-perspective analysis and spatiotemporal mapping of air pollution monitoring data, *Environmental Science and Technology* 44 (2010) 6738–6744.
- [23] A. Kolovos, J. M. Angulo, K. Modis, G. Papantonopoulos, J.-F. Wang, G. Christakos, Model-driven development of covariances for spatiotemporal environmental health assessment, *Environmental Monitoring and Assessment* 185 (2013) 815–831.
- [24] A. Kolovos, Spatiotemporal analysis of solar radiation for sustainable research in the presence of uncertain measurements, in: *Proceedings of the StatGIS 2009 Conference*, Milos, Greece, June 17–19, 2009.
- [25] S.-J. Lee, B. Bush, R. George, Analytic science for geospatial and temporal variability in renewable energy: A case study in estimating photovoltaic output in Arizona, *Solar Energy* 85 (2011) 1945–1956.
- [26] S. Salvador, P. Chan, Learning states and rules for detecting anomalies in time series, *Applied Intelligence* 23 (3) (2005) 241–255.
- [27] M. Halkidi, Y. Batistakis, M. Vazirgiannis, On clustering validation techniques, *Journal of Intelligent Information Systems* 17 (2-3) (2001) 107–145. doi:10.1023/A:1012801612483.
- [28] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Prez, I. Perona, An extensive comparative study of cluster validity indices, *Pattern Recognition* 46 (1) (2013) 243 – 256.
- [29] Solar Anywhere, *SolarAnywhere data*, Clean Power Research 2012 (2012).
URL <http://www.solaranywhere.com>
- [30] R. H. Inman, H. T. C. Pedro, C. F. M. Coimbra, Solar forecasting methods for renewable energy integration, *Prog. Energy Combust. Sci.* 39 (2013) 535 – 576.
- [31] P. Ineichen, R. Perez, A new airmass independent formulation for the linke turbidity coefficient, *Solar Energy* 73 (3) (2002) 151–157.
- [32] B. J. Frey, D. Dueck, Clustering by passing messages between data points, *Science* 315 (2007) 972–976.
- [33] K. Wang, J. Zhang, D. Li, X. Zhang, T. Guo, Adaptive affinity propagation clustering, *Acta Automatica Sinica* 33 (12) (2007) 1242–1246.
- [34] R. Steuer, J. Kurths, C. Daub, J. Weise, J. Selbig, The mutual information: Detecting and evaluating dependencies between variables, *Bioinformatics* 18 Suppl.2 (2002) S231–S240.
- [35] Y. S. Son, J. Baek, A modified correlation coefficient based similarity measure for clustering time-course gene expression data, *Pattern Recognition Letters* 29 (3) (2008) 232 – 242.
- [36] T. Caliński, J. Harabasz, A dendrite method for cluster analysis, *Communications in Statistics-theory and Methods* 3 (1) (1974) 1–27.
- [37] P. J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics* 20 (0) (1987) 53 – 65.
- [38] G. Christakos, *Random Field Models in Earth Sciences*, Academic Press, New York, NY, 1992.
- [39] P. Goovaerts, *Geostatistics for Natural Resources Evaluation*, Oxford University Press, New York, NY, 1997.
- [40] J. P. Chilés, P. Delfiner, *Geostatistics-Modeling Spatial Uncertainty*, John Wiley & Sons, New York, NY, 1999.
- [41] H. Wackernagel, *Multivariate Geostatistics: An Introduction with Applications*, 3rd Edition, Springer, Berlin, 2003.
- [42] G. Christakos, A bayesian/maximum-entropy view to the spatial estimation problem, *Mathematical Geology* 22 (7) (1990) 763–767.
- [43] G. Christakos, *Modern Spatiotemporal Geostatistics*, Oxford University Press, New York, NY, 2000.
- [44] G. Christakos, D. T. Hristopulos, *Spatiotemporal Environmental Health Modelling: A Tractatus Stochasticus*, Kluwer Academic Publishers, Boston, MA, 1998.
- [45] G. Christakos, P. Bogaert, M. L. Serre, *Temporal GIS*. With CD-ROM, Springer-Verlag, New York, NY, 2002.
- [46] Matlab Central, *Matlab Central File Exchange*, Adaptive Affinity Propagation Clustering s/w, Kaijun Wang (2009).
URL <http://www.mathworks.com/matlabcentral/fileexchange/18244-adaptive-affinity-propagation-clustering>
- [47] J. Astola, P. Haavisto, Y. Neuvo, Vector median filters, *Proceedings of the IEEE* 78 (4) (1990) 678–689.
- [48] A. Kolovos, G. Christakos, D. T. Hristopulos, M. L. Serre, On certain classes of non-separable spatiotemporal covariance models, *Advances in Water Resources* 27 (2004) 815–830.
- [49] A. Kolovos, A. Zagouras, L. Nonnenmacher, C. F. M. Coimbra, Methodology for deployment of optimally distributed solar monitoring networks, in: *Abstracts of the International Conference on Statistical Physics*, Statistical Physics 2014, Rhodes, Greece, July 7–11, 2014.
- [50] G. Christakos, D. T. Hristopulos, P. Bogaert, On the physical geometry concept at the basis of space/time geostatistical hydrology, *Advances in Water Resources* 23 (2000) 799–810.
- [51] R. A. Olea, *Geostatistics for Engineers and Earth Scientists*, Kluwer Academic, Boston, MA, 1999.

Figure 5: Step 2 predicted GHI for MS configurations A, B and C at 2 selected consecutive time instances on June 1, 2010. All plots in W/m^2 units; coordinates are measured in decimal degrees (DD). (a) Configuration A BME mean at time instance $t = 4$ (07.30); (b) Configuration A BME mean at time instance $t = 5$ (08.00); (c) Configuration B BME mean at time instance $t = 4$ (07.30); (d) Configuration B BME mean at time instance $t = 5$ (08.00); (e) Configuration C BME mean at time instance $t = 4$ (07.30); (f) Configuration C BME mean at time instance $t = 5$ (08.00).

Figure 6: Step 2 GHI prediction standard error for MS configuration A at selected time instances on June 1, 2010. All plots in W/m^2 units; coordinates are measured in decimal degrees (DD). (a) BME standard error at time instance $t = 4$ (07.30); (b) BME standard error at time instance $t = 5$ (08.00).

Figure 7: Distribution of relative error values for each MS configuration A, B, and C across all spatial locations at all instances in validation Step 2. The horizontal axis shows percentage of relative error, and the vertical axis is the corresponding count of validation grid points.