Automatic Parameter Optimization for Support Vector Regression for Land and Sea Surface Temperature Estimation From Remote Sensing Data

Gabriele Moser, *Member, IEEE*, and Sebastiano B. Serpico, *Senior Member, IEEE*

***Abstract*—Land surface temperature (LST) and sea surface temperature (SST) are important quantities for many environmental models. Remote sensing is a source of information for their estimation on both regional and global scales. Many algorithms have been devised to estimate LST and SST from satellite data, most of which require *a priori* information about the surface and the atmosphere. A recently proposed approach involves the use of support vector machines (SVMs). Based on satellite data and corresponding *in situ* measurements, they generate an approximation of the relation between them, which can subsequently be used to estimate unknown surface temperatures from additional satellite data. Such a strategy requires the user to set several internal parameters. In this paper, a method is proposed for automatically setting these parameters to quasi-optimal values in the sense of minimum estimation errors. This is achieved by minimizing a functional correlated to regression errors (i.e., the "span-bound" upper bound on the leave-one-out (LOO) error) which can be computed by using only the training set, without need for a further validation set. In order to minimize this functional, Powell's algorithm is adopted, since it is applicable also to nondifferentiable functions. Experimental results yielded by the proposed method are similar in accuracy to those achieved by cross-validation and by a grid search for the parameter configuration which yields the best test-set accuracy. However, the proposed method gives a dramatic reduction in the computational time required, particularly when many training samples are available.**

***Index Terms*—Generalization error bounds, land surface temperature (LST), Powell's method, sea surface temperature (SST), supervised regression, support vector machines (SVMs).**

I. Introduction

SATELLITE remote sensing systems provide a repetitive and consistent view of the Earth's surface at different observation scales. Measurements made by onboard sensors are a feasible way to derive Earth-surface parameters from surface-leaving radiation on both regional and global scales. Land surface temperature (LST) and sea surface temperature (SST) are important quantities for many environmental models, such as the energy and mass exchange between the atmosphere and surface, vegetation and soil moisture, numerical weather prediction, climatic variability, and global ocean circulation [1].

Many algorithms have been devised to obtain LST and SST from space radiometry [2]–[5]. Most of them require *a priori* information about the state of the atmosphere (e.g., temperature and water vapor profiles) and about the surface (e.g., emissivity), without which accurate surface temperature estimation becomes more difficult [6]–[12]. By itself, the estimation of these surface and atmospheric parameters is often a difficult task to be addressed prior to or jointly with surface temperature estimation [3], [8], [12]. Traditional methods, such as the well-known split-window techniques [1], are typically calibrated (once and for all) for given sensors (e.g., NOAA–AVHRR, MODIS, Meteosat Second Generation (MSG)–SEVIRI) by least squares fitting. These regressions are applied according to predefined sets of values of surface and atmospheric data and to the resulting simulated top-of-atmosphere brightness temperatures [5], [12], [13]. They can then be applied to a satellite image in an unsupervised way, i.e., with no need for training *in situ* temperature data, even though the aforementioned surface and atmospheric data for the related geographical area are still required. Thorough reviews of the LST and SST estimation problems from satellite data can be found in [1], [13], and [14].

A different approach to LST estimation has recently been proposed in [15] based on pattern recognition, in particular on support vector machines (SVMs). SVMs represent a family of powerful supervised learning techniques, and their application to the temperature estimation problem allows one to compute a nonparametric approximation of the relationship between satellite data and matching *in situ* measurements (adopted for training purposes). This supervised regression strategy can be theoretically proven to exhibit very good generalization and robustness properties [16]. It has also been experimentally demonstrated that such a strategy can generate more accurate estimates, as compared with the aforementioned standard techniques, even though at the expense of increased complexity and computational time [15]. In the context of remote sensing, SVM regression has recently been used also for biophysical parameter estimation from multispectral images of sea water [17]–[20], for evapotranspiration evaluation from meteorological, flux, and MODIS data [21], and for leaf-area-index estimation from radiometric data [22].

Unlike traditional methods, the SVM-based approach is supervised and requires *in situ* temperature measurements to be used for training purposes, but it does not involve an

Manuscript received September 25, 2007; revised February 21, 2008 and June 24, 2008. First published February 3, 2009; current version published February 19, 2009. This work was supported by the Italian Department of Civil Protection under the PROSCENIO 2005–2008 program.

The authors are with the Department of Biophysical and Electronic Engineering (DIBE), University of Genoa, 16145 Genova, Italy, and also with the Interuniversity Research Center in Environmental Monitoring (CIMA), 17100 Savona, Italy (e-mail: gabriele.moser@unige.it; sebastiano.serpico@unige.it).

Digital Object Identifier 10.1109/TGRS.2008.2005993

explicit characterization of surface and/or atmospheric parameters. From this perspective, such an approach is complementary with respect to the traditional methods. Radiometric and/or subsurface ground temperature measurements are regularly acquired by several networks of micrometeorological stations (e.g., the well-known "Oklahoma Mesonet" network in Oklahoma (USA) or similar networks installed in several river basins for flood prevention and environmental monitoring). Sea temperature measurements are acquired on a regular basis by suitably equipped moored and drifting buoys.

The SVM-based strategy in [15] involves tuning several internal parameters, whose values influence the SVM-generated approximation function and must be preliminarily set by the user, typically via "trial-and-error" grid searches for the lowest regression error (e.g., cross-validation (CV) or LOO error). This process is often time consuming, due to the need to repeat the SVM training for each node in the grid, and it is also sensitive to the (usually empirical) choice of the related ranges and discretization steps of the parameters. In particular, the time requirement for grid searches can become critical when large *in situ* data sets are available for training purposes, which may occur, for instance, when *in situ* stations continuously provide temperature measurements over a long time. This calls for the availability of fast and automatic parameter-setting procedures.

In this paper, we focus on the problem of the automatic optimization of the input parameters characterizing an SVM regression method. In the literature, this problem has been mainly addressed in the context of support vector classification and has been expressed as the numerical minimization of upper bounds on the generalization error of the classifier [23], [24]. In [23], several of such bounds (e.g., the "radius-margin bound" and the "span bound") are introduced or revised while focusing on the family of support vector classifiers with quadratic penalty on the slack variables [16]. Gradientlike minimization procedures are also developed for such bounds, provided that they are differentiable. This approach is extended in [24] to support vector classifiers with linear penalty on the slack variables [16] by combining a heuristic generalization of the radius-margin bound and a quasi-Newton minimization algorithm. In [25], several error bounds (i.e., the "$\xi\alpha$ bound" [26], the "generalized approximate CV" [27], the "approximate span bound" [28], the "Vapnik–Chervonenkis bound" [29], and the aforementioned radius-margin bound) are revised and experimentally compared in the contexts of support vector classification with both linear and quadratic penalties on the slack variables. In [30], quantitative bounds on the error probability of general families of discriminant functions are formulated in terms of classification errors on the training samples and of an analytical measure of the complexity of the classifier (i.e., the Rademacher complexity).

As far as support vector regression is concerned, semiheuristic approaches to parameter setting have been proposed in [31] and [32] by relating the parameters with the noise and input data statistics. Bayesian methods for parameter estimation have been proposed in [33] and [34], even though they focus on case-specific SVM architectures (without the bias term in [34] and with an *ad hoc* loss function in [33]). In [35], the approach based on gradientlike minimization of error bounds is extended to support vector regression: Two parameter optimization methods are developed by generalizing to regression the aforementioned radius-margin and span bounds and by combining them with quasi-Newton procedures while focusing on SVM with quadratic penalty on the slack variables. Note that, in this approach, the radius-margin bound for regression exhibits quite a low correlation with actual regression errors on test samples and that the span bound for regression is a nondifferentiable function of the unknown parameters. To overcome the latter issue, a differentiable modified bound is used in [35], even though this requires the introduction of a further smoothing parameter to be manually set. The Rademacher complexity bound holds in the case of regression as well [30], [36], and a closed-form analytical formulation of the bound is available for a kernel-based functional approximator (such as the ones given by SVMs), even though it does not have a bias term.

In this paper, a method is proposed that automatically sets the parameters required by SVM regression in order to minimize estimation errors. The method is applied in the context of the SVM approach to LST estimation; this approach is also extended to the SST case and can be generalized to the estimation of further bio-/geophysical parameters as well. The proposed automatic parameter optimization method lies in minimizing the span-bound functional, which is known to be strongly correlated with estimation errors and can be computed by using only the training set [35]. Unlike [35], the SVM with linear penalty on the slack variables is adopted, since it usually ensures better sparsity of the resulting functional approximator, while limiting possible risks of overfitting and reducing the time required to generate maps of LST and SST estimates, when compared to using a quadratic penalty [16]. Powell's method is the algorithm chosen to address the minimization problem because it is applicable to nondifferentiable functions as well [37], such as the considered span bound. This prevents the need to introduce regularized versions of the functional, which also involve further parameters to be manually set.

The main novelty of this paper lies in the combination of the span-bound functional for SVM regression with linear penalty on the slack variables and Powell's algorithm, in order to obtain complete automation of the estimation of LST and SST from satellite data. Here, we do not focus on the applicative aspects related to LST and SST retrieval from satellite data, but we rather concentrate on the methodological problem of automatic parameter optimization for support vector regression. Accordingly, the experimental results regarding LST estimation from AVHRR and MSG images and SST estimation from AVHRR data are compared with those obtained by a frequently applied approach to parameter setting, i.e., the grid search for the parameter configuration yielding the smallest CV or test-set (hold-out) error. More application-oriented comparisons with classical LST and SST estimation algorithms can be found in [15] and [38].

This paper is organized as follows. The proposed approach is described in Section II, after briefly revising the main ideas and notations of support vector regression. Experimental results are presented in Section III, and conclusions are drawn in Section IV.

II. Methodology

A. Support Vector Regression for LST and SST Estimation

Postulating the existence of a relation between the surface temperature of an observed geographical area and the spectral radiance measured by a remote sensor over this area, an evaluation of the surface temperature would require the inversion of this relation: From a learning-oriented perspective, this suggests using regression methods. Accordingly, a supervised estimation scheme based on SVMs has been adopted and applied to passive sensor data in [15].

The SVM approach aims at expressing a learning task (i.e., classification, regression, or probability density estimation [39]) by explicitly optimizing a measure of the optimization capability of the learning machine. Given a training set for which the corresponding *in situ* measurements are known, the SVM generates an approximation of the relation between satellite data acquired over a certain area and the corresponding LST.

Specifically, denoting x as the n-dimensional feature vector extracted from the acquired sensor data for a given pixel, y as the corresponding Earth-surface physical quantity to be estimated (either LST or SST), and $f : \mathbb{R}^n \to \mathbb{R}$ as the unknown function relating the features with this physical quantity (i.e., $y = f(x)$, up to possible noise), the SVM approach allows an estimate $\hat{f}$ by minimizing an upper bound on the probability that the estimation error may be above a given threshold [16]. Given a set $\{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ of N training samples (where x_h is the feature vector for the hth training sample and y_h is the corresponding *in situ* measurement; $h = 1, 2, \ldots, N$), the resulting approximation turns out to be expressed as a linear combination of appropriate kernel functions centered on a subset of training samples [16]

$$\hat{f}(x) = \sum_{h \in \mathcal{S}} \beta_h^* K(x_h, x|\gamma) + b^* \tag{1}$$

where $\beta_1^*, \beta_2^*, \ldots, \beta_N^*$ are the weight coefficients of the linear combination, $K(\cdot, \cdot|\gamma)$ is a kernel function parameterized, in general, by a vector γ of r real-valued parameters ($\gamma \in \mathbb{R}^r$), b^* is a bias term, and $\mathcal{S} = \{h : \beta_h^* \neq 0\}$ [16]. If $h \in \mathcal{S}$, i.e., $\beta_h^* \neq 0$, the training sample x_h is named "support vector." Necessary and sufficient conditions (the so-called "Mercer's conditions") are known for a given function of two vectors to be a kernel [16]. For instance, Gaussian and symmetric hyperbolic tangent functions are popular kernels. We assume that the vector γ of the internal parameters of the kernel is defined such that it can take on values in the whole space $\mathbb{R}^r$. The vector[1] $\beta^* = [\beta_1^*, \beta_2^*, \ldots, \beta_N^*]^\top$ of the coefficients characterizing the kernel expansion in (1) is obtained by solving the following quadratic programming (QP) problem [16]:

$$\begin{cases} \min\limits_{\beta \in \mathbb{R}^N} \left[\frac{1}{2}\beta^\top Q(\gamma)\beta - y^\top \beta + \varepsilon\|\beta\|_1\right] \\ \mathbf{1}^\top \beta = 0, \quad \|\beta\|_\infty \leq C \end{cases} \tag{2}$$

where C and ε are further SVM parameters, $Q(\gamma)$ is the $N \times N$ matrix whose (h, k)-entry is given by $Q_{hk}(\gamma) = K(x_h, x_k|\gamma)$ ($h, k = 1, 2, \ldots, N$), $y = [y_1, y_2, \ldots, y_N]^\top$ is the vector of the N *in situ* measurements, and $\mathbf{1}$ is an N-dimensional vector with unitary components. An important property of SVMs is that the resulting QP solution is globally optimal, while other supervised learning techniques (e.g., neural networks) only ensure a local minimum of the approximation error to be found [16]. Numerical techniques to efficiently solve this QP problem are available in the literature [40], [41].

The parameter $C (C > 0)$ tunes the tradeoff between the generalization capability of the functional approximator and the accuracy of the fitting on the training set: A large value of C allows a high regression accuracy on the training set to be achieved, but may encourage overfitting, thus causing a high sensitivity to noise and to possible errors on the training data. A small value of C aims at optimizing the smoothness and generalization capability of the approximator, but may poorly fit the training set, thus causing a reduced accuracy of the approximation of the estimated function. The parameter $\varepsilon (\varepsilon > 0)$ represents the largest (absolute) error that is considered acceptable and is not penalized at all in the SVM training process [16]. The vector of all the input parameters of the SVM regression method can be defined as $\theta = (\ln C, \ln \varepsilon, \gamma) \in \mathbb{R}^m$, where $m = r + 2$ is the number of parameters, and the logarithm is applied to C and ε in order to let θ take on values in the whole space $\mathbb{R}^m$, with no explicit positivity constraint. Given a value for θ, the training of the SVM consists in the solution of the QP problem in (2): We stress the dependence on θ by denoting as $\beta^*(\theta)$ and $\mathcal{S}(\theta)$ the resulting weight vector of the linear combination and the related index set of the support vectors, respectively. A more detailed discussion of SVM regression can be found in [16].

B. Proposed Parameter Optimization Method

The key idea of the proposed parameter optimization technique is to search for a parameter vector that minimizes an upper bound on the regression error. Specifically, one can prove that, under mild assumptions, the average "LOO" regression error can be upper bounded by the following functional, named "span bound" [35]:

$$J(\theta) = \frac{1}{N} \sum_{h \in \mathcal{S}(\theta)} |\beta_h^*(\theta)|\, s_h^2(\theta) + \frac{y^\top \beta^*(\theta) - \varepsilon \|\beta^*(\theta)\|_1 - \beta^*(\theta)^\top Q(\gamma)\beta^*(\theta)}{CN} + \varepsilon \tag{3}$$

where $s_h(\theta)$, named the "span" of the support vector x_h, is a real coefficient computed by solving the following QP

[1]Given that $p \geq 1$ and $v \in \mathbb{R}^m$, we denote by $\|v\|_p$ and $\|v\|_\infty$ the pth-order Minkowski norm (i.e., $\|v\|_p = (|v_1|^p + |v_2|^p + \cdots + |v_m|^p)^{1/p}$), and the Tchebitchev norm (i.e., $\|v\|_\infty = \max\{|v_1|, |v_2|, \ldots, |v_m|\}$) of v, respectively. All vectors in this paper are assumed to be column vectors, and the superscript "$\top$" denotes the matrix transpose operator.

problem ($h \in \mathcal{S}(\theta)$):

$$s_h^2(\theta) = \min \left\{ \sum_{i,j \in \mathcal{U}(\theta)} \lambda_i \lambda_j Q_{ij}(\gamma) - 2 \sum_{i \in \mathcal{U}(\theta)} \lambda_i Q_{ih}(\gamma) + Q_{hh}(\gamma) \middle| \lambda \in \mathbb{R}^N, \sum_{i \in \mathcal{U}(\theta)} \lambda_i = 1 \right\} \quad (4)$$

and $\mathcal{U}(\theta) = \{h \in \mathcal{S}(\theta) : |\beta_h^*(\theta)| < C\} = \{h : 0 < |\beta_h^*(\theta)| < C\}$ identifies a subset of the training set, whose samples (i.e., x_h such that $h \in \mathcal{U}(\theta)$) are named "unbounded support vectors" or "free support vectors".

Given the parameter vector θ, the corresponding value $J(\theta)$ of the span bound can be analytically computed by solving first the QP problem in (2), which yields the coefficients $\beta_h^*(\theta)(h = 1, 2, \ldots, N)$ and identifies support vectors and unbounded support vectors, and then the QP problem in (4) for each support vector $x_h (h \in \mathcal{S}(\theta))$.

The proposed method aims at searching a parameter vector $\theta^* \in \mathbb{R}^m$ that minimizes $J(\theta)$. Even though the span bound can be computed by dealing only with the training samples, it is known to be a very tight bound, strongly correlated with the regression error on test samples, provided that the test and training samples are drawn from the same probability distribution [35]. However, in general, it is a nondifferentiable function of θ, so gradient-based algorithms cannot be used to minimize it. Methods that only need to evaluate the functional to be minimized are required. In particular, Powell's method is adopted here, which is an iterative unconstrained minimization technique evaluating only the functional and not requiring its possible derivatives. Powell's method aims at emulating, without the use of gradient and Hessian information, the behavior of the conjugate gradient method (which, by itself, is applicable to twice continuously differentiable functions) by identifying a set of m vectors in $\mathbb{R}^m$ that are close to being conjugate. The minimum of the functional is computed by sequentially performing, at each iteration, a line search[2] in each of such "pseudoconjugate" directions [42].

Let θ^t be the parameter vector computed at the tth iteration of the proposed method, $u^{i,t}$ be the ith pseudoconjugate direction computed at the tth iteration, and $\bar{\theta}^{i,t}$ be the parameter vector computed after the ith sequential line search of the tth iteration ($t = 0, 1, 2, \ldots$ and $i = 1, 2, \ldots, m$).

The method is initialized with a given parameter vector θ^0 and with the assumption that the initial directions $u^{1,0}, u^{2,0}, \ldots, u^{m,0}$ are the canonical basis vectors in $\mathbb{R}^m$ (i.e., $u^{i,0}$ has a unitary ith component, while all the other components are zero; $i = 1, 2, \ldots, m$). Then, at each tth iteration of the proposed technique, the following operations are performed ($t = 0, 1, 2, \ldots$) [37], [42].

1) Set $\bar{\theta}^{0,t} = \theta^t$.
2) For each ith pseudoconjugate direction $u^{i,t}$, start from the current point $\bar{\theta}^{i-1,t}$, and minimize $J(\theta)$ by a line search in the direction $u^{i,t}$, i.e., compute the scalar
$$\rho^{i,t} = \arg\min_{\rho \in \mathbb{R}} J(\bar{\theta}^{i-1,t} + \rho u^{i,t}) \quad (5)$$
and then move to the resulting minimum point $\bar{\theta}^{i,t} = \bar{\theta}^{i-1,t} + \rho^{i,t} u^{i,t} (i = 1, 2, \ldots, m)$.
3) Update the parameter vector as the last point reached in the parameter space by the sequence of m line searches performed in step 2), i.e., set $\theta^{t+1} = \bar{\theta}^{m,t}$.
4) Update the search directions by dropping the first direction $u^{1,t}$ and introducing a new direction $(\theta^{t+1} - \theta^t)$, i.e., set $u^{i,t+1} = u^{i+1,t}$ for $i = 1, 2, \ldots, m-1$ and $u^{m,t+1} = \theta^{t+1} - \theta^t$.

These processing steps are iterated until the relative decrease in the span-bound functional between successive iterations goes below a given threshold, which is very close to the machine precision [37]. The method is analytically proved to quadratically converge, under mild assumptions, at least to a local minimum [37], [42]. A drawback is that the pseudoconjugate directions may become linearly dependent during the iterations, thus implicitly restricting the search space to a lower dimensional subspace [43]. A simple strategy to solve this problem while keeping the quadratic convergence property consists of reinitializing the pseudoconjugate directions to the canonical basis of $\mathbb{R}^m$ after each cycle of $(m + 1)$ successive iterations [37]. Alternate solutions to this problem, either keeping or giving up the quadratic convergence property, are discussed in [37] and [43].

The line search in step 2) is performed numerically using Brent's method, which is based on the iterative combination of the "inverse parabolic interpolation" and "golden-section" approaches to single-variable function minimization and is applicable to nondifferentiable functions as well [43]. Given a single-variable continuous function to be minimized, the golden-section method iteratively computes a sequence of bracketing triplets[3] that converges to a local minimum point of the function [37]. The inverse parabolic interpolation method iteratively minimizes a parabolic interpolation of the input (single-variable) function computed according to the evaluations of the function in three distinct points. If the function is monomodal, this method is proved to converge to the desired local minimum point. Moreover, it typically converges faster than the golden-section approach. However, if the monomodality assumption fails, it may not converge [37]. Brent's method is based on the key idea of iteratively exploiting the inverse parabolic interpolation approach when close to a monomodal valley of the function to be minimized while using the golden-section technique otherwise. Basically, during each iteration, a parabolic interpolation step is attempted, and a series of tests is performed to discriminate if the resulting solution is acceptable or if a golden-section step is to be applied. Further details about the resulting numerical procedure, which is not straightforward,

[2]By "line search," we mean the search for the minimum of a given function of m variables along a straight line in the related m-dimensional space.

[3]We recall that, given a continuous function $g : \mathbb{R} \to \mathbb{R}$, a "bracketing triplet" is a triplet $(a, b, c) \in \mathbb{R}^3$ such that $a < b < c$ and $g(b) \leq \min\{g(a), g(c)\}$. Thanks to the continuity of g, this condition ensures that a local minimum point of g exists in (a, c).

can be found in [43] and [37]. The proposed "Powell span-bound" technique will be denoted by PSB in the following.

III. Experimental Results

A. Data Sets and Experimental Setup

The method was tested on three real data sets, with two of them being related to the LST estimation problem and one being related to the SST case. The satellite data of the first LST data set (named "LST-AVHRR-Mesonet") were 14 AVHRR cloud-free images collected around 7:30 A.M. (local time) over Oklahoma (USA) between March and September 2004. The related ground observations are *in situ* measurements of subsurface soil temperatures at 5 cm under native vegetation acquired every 15 min by "Oklahoma Mesonet," an automated network of surface micrometeorological stations covering Oklahoma (for each image, the ground measurements obtained for the nearest quarter of an hour were considered). The resulting data set was randomly split into 414 training samples and 375 testing samples. Note that, while the available *in situ* data were subsurface point measurements taken at 5 cm, the satellite brightness temperatures are related to the overall radiative surface temperature of the area corresponding to each pixel. In general, these two data typologies may differ from each other. However, as discussed in [44], the difference between the 5-cm subsurface temperature and the skin temperature obtained by correcting infrared radiative temperature measurements for emissivity is very small around 7:00–8:00 A.M. (local time), which is the acquisition time of the considered AVHRR images. More generally, SVM regression does not attempt to interpolate the training *in situ* samples, but it computes a functional approximator of the relationship between satellite and *in situ* data. Thanks to the well-known generalization properties of SVMs, possible small differences between subsurface and radiative temperatures at the acquisition time are expected not to significantly affect regression accuracy.

To apply SVM, three features were considered for each sample, namely, the measured brightness temperatures of AVHRR channel 4, the difference between the brightness temperatures of channels 4 and 5, and the "normalized difference vegetation index" (NDVI). One of the two brightness temperatures and the difference between them were used instead of the two brightness temperatures *per se* because the difference feature is known to be important for temperature estimation purposes, thanks to its correlation with the atmospheric effects on sensor radiance [45]. The choice of these features extracted from AVHRR data was suggested by their use in popular physically based split-window techniques proposed for this sensor [46]–[48].

The second LST data set (named "LST-MSG-ARSIA") is composed of images acquired over Italy between 7:00 and 9:00 A.M. (local time) by the SEVIRI instrument onboard (MSG) between August 12 and September 19, 2005, and received by the station of the Interuniversity Research Center in Environmental Monitoring (CIMA) in Savona (Italy). The related *in situ* data are radiometric ground temperature measurements acquired by 13 micrometeorological stations of the "Agenzia Regionale per lo Sviluppo e l'Innovazione nel Settore Agricolo e Forestale" (ARSIA) network of the Italian Tuscany region. Samples affected by cloud cover were removed from the data set. The samples acquired by 7 out of the 13 available stations were used for training purposes; the ones acquired by the other 6 stations were employed for testing. The resulting 149 training and 104 test samples were spatially disjoint. The employed features were the brightness temperature of the 10.82-μm SEVIRI channel, the difference between the brightness temperatures in the 10.82- and 12-μm channels, and the NDVI feature.

For the SST case, a data set (named "SST") was constructed from the publicly available multiyear multisatellite "AVHRR Pathfinder Matchup Database" of approximately cotemporal colocated *in situ* SSTs (acquired from both moored and drifting buoys) and AVHRR images. Details on this database can be found in [49]. Samples collected on two days, i.e., January 1 and 2, 1999, from globally distributed buoys were used for experiments. Out of the 215 available samples, 109 were randomly chosen for training purposes, and the remaining 106 were used for testing. As in [49] and [50], four features were considered for each sample, i.e., the secant of the satellite zenith angle, the measured brightness temperatures of AVHRR channels 4 and 3B, and the difference between the brightness temperatures of channels 4 and 5. The channel-3B (midwave infrared) brightness temperature was used for its correlation with nighttime SST [50] and because the samples in these data set were collected during two whole days, including both daytime and nighttime acquisitions. In the daytime, channel 3B is also affected by solar radiance. Thus, from the viewpoint of SST estimation, this feature is expected to be more noisy in samples collected in the daytime than in the nighttime. Anyway, this fact will not be critical due to both the high correlation coefficient with SST *in situ* measurements (around 89%, irrespective of acquisition time) and the robustness of SVM to noise.

For all three data sets, all the features of the training and test samples and the temperature values were normalized to the interval [0, 1], and SVM was applied with a Gaussian radial basis function kernel, i.e. $(u, v \in \mathbb{R}^n)$

$$K(u, v|\sigma) = \exp\left(-\frac{\|u - v\|^2}{2\sigma^2}\right)$$

which is parameterized by the standard deviation $\sigma(\sigma > 0)$. This kernel has frequently been employed in remote sensing applications, and the experimental analysis described in [15] confirms that it allows accurate LST estimates to be performed. The resulting parameter vector is defined as $\theta = (\ln C, \ln \varepsilon, \ln \sigma)$, so that all three components of θ can range in the whole real line.

B. Experiment I: Correlation Between Span Bound and Test-Set Errors

A first experiment was performed in order to preliminarily investigate the choice of the span bound as a functional guiding the parameter optimization process. Focusing on

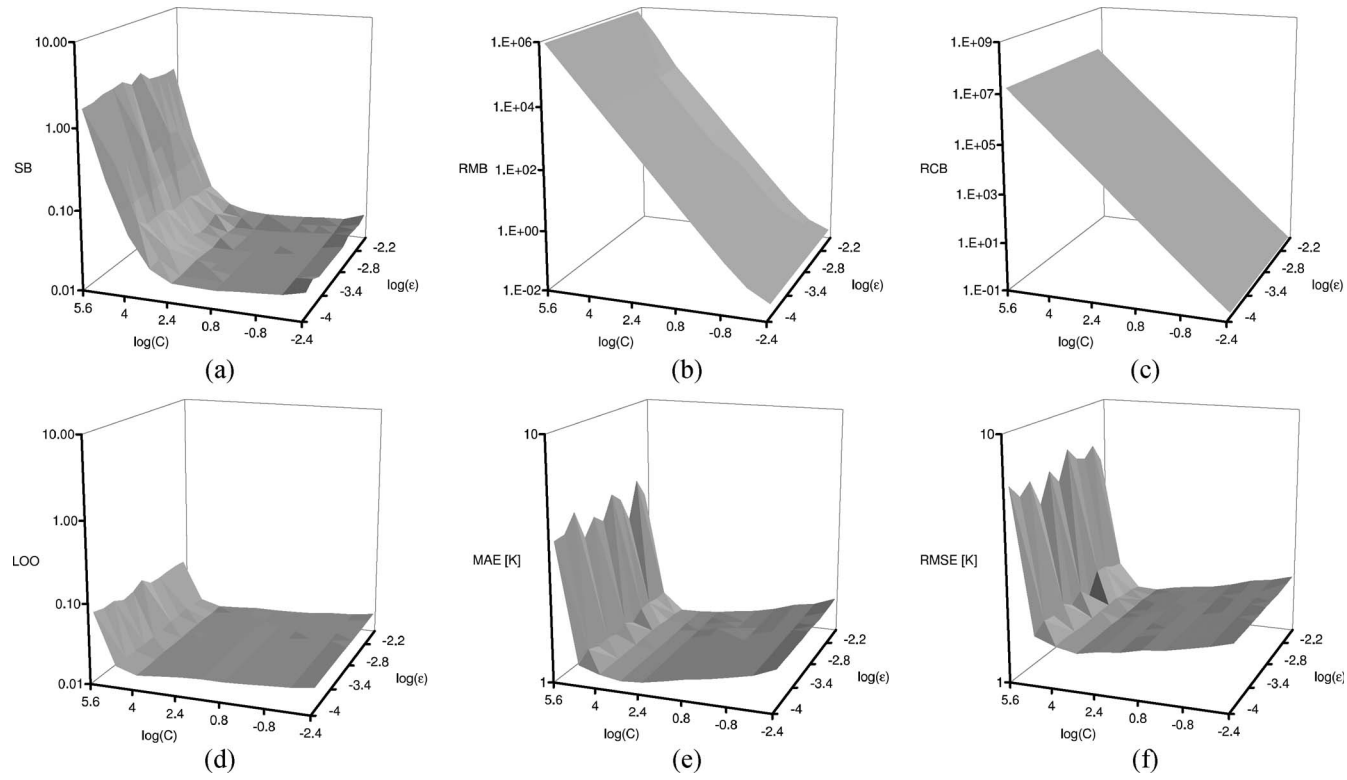

Fig. 1. Experiment I: Plots of (a) the span bound, (b) the radius-margin bound, (c) the Rademacher complexity bound, and (d) the LOO error (denoted by SB, RMB, RCB, and LOO, respectively, and computed on the training set) and of (e) MAE and (f) rmse (in [K]), computed on the test set as functions of the parameters C and ε (with $\sigma = 0.25$).

"LST-AVHRR-Mesonet," the span bound, the radius-margin bound, the Rademacher complexity bound (computed according to the training samples), the LOO regression error, the mean absolute error (MAE), and the root-mean-square error (rmse) on the test set were computed as functions of C and ε, with $C \in [10^{-3}, 10^{6}]$ and $\varepsilon \in [10^{-4}, 10^{-2}]$ (the kernel parameter σ was set to a constant value, i.e., 0.25, so that the plots could be drawn; see Fig. 1).

In the analyzed ranges of C and ε, the span bound, the LOO error, and the test MAE and rmse have similar valleylike nearly ε-insensitive shapes, with small and large values of C causing larger function values. This is an expected behavior, according to the meaning of the parameter C in SVM training (see Section II-A). However, MAE, rmse, and the LOO error have slightly wider valley floors than the span bound; this suggests that the span bound may tend to overestimate regression errors for large values of C. In particular, the span bound is a very tight upper bound on the LOO error in the valley floor area (on average, the span bound is just 0.8% higher than the LOO error in this area), while it is a coarser bound for large values of C. Although not shown here for brevity, the behaviors of MAE and rmse on the training set are very similar to those of their test counterparts, except for slightly better overall values (as expected).

On the contrary, both the radius-margin and Rademacher complexity bounds exhibited a strictly monotonic behavior with no local/global minima, which makes them unfit to guide the parameter optimization process, at least on the considered data set. The radius-margin bound is known to be a loose bound on the LOO error and to often exhibit a low correlation with test-set errors [35]. With regard to the Rademacher complexity, it is worth noting that the related bound holds for a kernel-based functional approximator with no bias term [30], whereas the SVM estimator in (1) includes the bias term b^*. Hence, the application of this error bound in the present case is not formally rigorous, which may explain the lack of correlation with the test-set errors. From this perspective, possible generalizations of the Rademacher complexity bound to kernel-based expansions with bias terms may be proven and tested as well, even though this further development would be outside the scope of this paper.

It is also worth noting that, even if the span bound and test MAE and rmse minima are not achieved by the same parameter configuration, the values of MAE and rmse obtained by the parameter configuration yielding the smallest span bound differ from the minima MAE and rmse by just 0.08 K and 0.13 K, respectively. The similar behaviors of these three functions of (C, ε) are also confirmed by the fact that the correlation coefficients between the span bound and MAE and between the span bound and rmse are 96.3% and 96.9%, respectively. This is pivotal, as it confirms that, by minimizing the span bound, the resulting parameter configuration may also lead to good regression performance on unknown samples. Similar results (not shown for brevity) were obtained for the behaviors of the considered functionals as functions of (C, σ) and (ε, σ) and with the other two data sets as well.

C. Experiment II: Temperature Estimation

The proposed PSB method was initialized with $C^0 = 1, \varepsilon^0 = 0.01$, and $\sigma^0 = 0.5$, and applied to each of the three considered data sets. Four and five iterations were sufficient for PSB to converge in the cases of "SST" and of the two LST data sets, respectively. The results were compared with the ones obtained by two grid-search approaches. In the first approach, a hold-out strategy was used, i.e., for each node in the grid, an SVM was trained on the training set, and then, the parameter vector yielding the smallest MAE on the test set was searched for in the grid. This represents an "ideal case" in which the true *in situ* temperatures are assumed to be known also on the test samples; this is not a realistic parameter optimization strategy but is included here as a benchmark experiment. The second grid approach searched for the parameter vector yielding the smallest CV MAE, computed, as usual, on the training samples; this is a feasible and popular parameter optimization technique. A three-fold CV, with a random subdivision of the training set into three disjoint almost-equal-sized subsets, was used. In both cases, the considered set of values were $C \in \{10^{-3}, 10^{-2}, \ldots, 10^{4}\}$, $\varepsilon \in \{10^{-4}, 10^{-3}, \ldots, 10^{-1}\}$, and $\sigma \in \{0.01, 0.02, \ldots, 1\}$.

The test-set accuracies obtained by using SVMs with the parameter configurations computed by PSB and by the two grid searches are presented in Tables I–III. All considered parameter optimization techniques allowed accurate temperature estimates to be generated with MAE $\simeq$ 1–1.5 K and rmse $\simeq$ 1.3–2 K in the LST case and MAE $\simeq 0.5$ K and rmse $\simeq 0.7$ K in the SST case. The correlation coefficient ρ between the true and estimated temperatures on the test samples was very high for the "SST" ($\rho \simeq 99\%$) and "LST-AVHRR-Mesonet" ($\rho > 97\%$) data sets and quite high for "LST-MSG-ARSIA" ($\rho \simeq 86\%$–88%). This is consistent with the fact that the estimated temperatures exhibited smaller mean errors with respect to the true ones in the cases of "LST-AVHRR-Mesonet" (mean error below 0.1 K) and "SST" (mean error below 0.2 K) than in the case of "LST-MSG-ARSIA" (mean error around 0.7 K). Similarly, the intercepts of the linear regression of the estimated temperatures with respect to the true ones were larger with "LST-MSG-ARSIA" than with "LST-AVHRR-Mesonet" and "SST." This is interpreted as a consequence of the lower correlation of the satellite infrared observations with the *in situ* measurements for "LST-MSG-ARSIA." For instance, the correlation coefficients between the *in situ* temperatures and the brightness temperature channels were 95% $\div$ 96% for both "SST" and "LST-AVHRR-Mesonet" and just 76% for "LST-MSG-ARSIA."

The parameter vectors identified by PSB were quite different from the grid-search ones. However, very similar performances were exhibited by the three approaches, with slightly smaller regression errors on the hold-out grid search. This is an expected result as this approach also exploits the true *in situ* temperatures of the test samples. However, the differences between the values of MAE for PSB and for the hold-out grid search were just 0.07 K for "LST-AVHRR-Mesonet" and "SST" and 0.11 K for "LST-MSG-ARSIA," and the differences between the corresponding rmse values were 0.07 K for "LST-MSG-ARSIA" and "SST" and 0.12 K for "LST-MSG-ARSIA."

TABLE I
EXPERIMENT II, DATA SET "LST-AVHRR-MESONET": MAE, RMSE, AND MEAN ERROR ON THE TEST SET FOR THE SVM FUNCTIONAL APPROXIMATORS OBTAINED BY THE PSB AND GRID-SEARCH PARAMETER OPTIMIZATION METHODS (CV, HOLD-OUT), CORRELATION COEFFICIENT ρ BETWEEN THE TRUE AND ESTIMATED TEMPERATURES OF THE TEST SAMPLES, SLOPE AND INTERCEPT OF THE LINEAR REGRESSION OF THE ESTIMATED TEMPERATURES WITH RESPECT TO THE TRUE ONES, COMPUTATION TIMES FOR THE TRAINING PHASES, AND TRAINING AND TEST-SAMPLE SIZES. THE RELATED PARAMETER VALUES ARE ALSO REPORTED

	PSB	CV	Hold-out
MAE [K]	1.10	1.05	1.03
RMSE [K]	1.46	1.35	1.34
ρ	97.06%	97.47%	97.51%
mean error [K]	0.07	-0.02	-0.04
slope	0.97	0.98	0.97
intercept [K]	0.52	0.47	0.63
time	36m 3s	6h 56m 33s	5h 52m 22s
training samples	414		
test samples	375		
C	122.635	10	10
ε	0.004	0.001	0.01
σ	0.244	0.08	0.08

TABLE II
EXPERIMENT II, DATA SET "LST-MSG-ARSIA": MAE, RMSE, AND MEAN ERROR ON THE TEST SET FOR THE SVM FUNCTIONAL APPROXIMATORS OBTAINED BY THE PSB AND GRID-SEARCH PARAMETER OPTIMIZATION METHODS (CV, HOLD-OUT), CORRELATION COEFFICIENT ρ BETWEEN THE TRUE AND ESTIMATED TEMPERATURES OF THE TEST SAMPLES, SLOPE AND INTERCEPT OF THE LINEAR REGRESSION OF THE ESTIMATED TEMPERATURES WITH RESPECT TO THE TRUE ONES, COMPUTATION TIMES FOR THE TRAINING PHASES, AND TRAINING AND TEST-SAMPLE SIZES. THE RELATED PARAMETER VALUES ARE ALSO REPORTED

	PSB	CV	Hold-out
MAE [K]	1.51	1.53	1.40
RMSE [K]	1.93	1.96	1.86
ρ	87.17%	86.83%	88.08%
mean error [K]	0.76	0.78	0.60
slope	0.72	0.71	0.67
intercept [K]	4.58	4.67	5.60
time	47s	2h 22m 33s	1h 55m 30s
training samples	149		
test samples	104		
C	4.976	10	0.1
ε	0.002	0.001	0.001
σ	0.491	0.46	0.42

Even smaller differences can be noted between the regression errors of PSB and of the CV grid search: Slightly smaller values of MAE and rmse were obtained by CV, when applied to "SST" and "LST-AVHRR-Mesonet," and by PSB, when applied to "LST-MSG-ARSIA." The fact that quite different parameter vectors computed by PSB and by the grid searches resulted in very similar regression performances can be explained by the valleylike behaviors of the span-bound, MAE, and rmse functionals described in Section III-B.

On the other hand, dramatic differences can be noticed between the computation times of PSB and of the grid-search for LST estimation (all times in Tables I–III refer to the same hardware configuration, i.e., a 3.4-GHz CPU with 1-GB RAM). The grid search, when applied to the two LST data sets in both the CV and hold-out versions, required some hours to identify optimal parameter vectors and to train the related SVM approximators due to the need to repeat the QP solution process

TABLE III
EXPERIMENT II, DATA SET "SST": MAE, RMSE, AND MEAN ERROR ON THE TEST SET FOR THE SVM FUNCTIONAL APPROXIMATORS OBTAINED BY THE PSB AND GRID-SEARCH PARAMETER OPTIMIZATION METHODS (CV, HOLD-OUT), CORRELATION COEFFICIENT ρ BETWEEN THE TRUE AND ESTIMATED TEMPERATURES OF THE TEST SAMPLES, SLOPE AND INTERCEPT OF THE LINEAR REGRESSION OF THE ESTIMATED TEMPERATURES WITH RESPECT TO THE TRUE ONES, COMPUTATION TIMES FOR THE TRAINING PHASES, AND TRAINING AND TEST-SAMPLE SIZES. THE RELATED PARAMETER VALUES ARE ALSO REPORTED

	PSB	CV	Hold-out
MAE [K]	0.54	0.46	0.46
RMSE [K]	0.73	0.67	0.67
ρ	99.43%	99.55%	99.55%
mean error [K]	0.14	0.19	0.19
slope	0.98	1.01	1.01
intercept [K]	0.36	-0.32	-0.32
time	2m 16s	4m 21s	4m 21s
training samples	109		
test samples	106		
C	3.457	10	10
ε	0.002	0.01	0.01
σ	0.500	1	1

at each node in the grid. The difference in computation times is less significant in the case of "SST" because the number of available training samples is smaller for this data set than for the two LST data sets, and the time required by each QP numerical solution is correspondingly much shorter [16]. As expected, the 3-fold CV turned out to be the most time-consuming technique, as it performed three QP solutions for each node in the grid, even though on smaller training sets as compared with the hold-out case. On the other hand, in order to complete the parameter optimization and training processes, PSB required about just 47 s, 2 min, and 36 min for "LST-MSG-ARSIA," "SST," and "LST-AVHRR-Mesonet," respectively. This is explained by the fact that grid searches perform a complete exploration of the parameter space (up to the adopted search ranges and quantization steps), whereas PSB performs a sequence of moves (in the parameter space), which approaches a locally optimum solution. From this viewpoint, PSB is a suboptimal approach, even though it still identifies parameter configurations affected by regression errors very close to the ones incurred by grid searches.

Fig. 2(b) and (c) shows the maps of LST estimates obtained by applying the SVMs trained on the "LST-MSG-ARSIA" training set (with the parameter vectors computed by PSB and by the hold-out grid search, respectively) to an MSG image acquired over Italy on September 15, 1999, at 7:00 A.M. local time [Fig. 2(a)]. A visual comparison between such maps further confirms the close similarity between the PSB and grid-search results. Similar remarks hold for the "LST-AVHRR-Mesonet" data set as well. The same visual conclusions can be drawn from Fig. 2(e) and (f), which shows the maps of SST estimates obtained by applying the SVMs trained on the "SST" training set (with the parameter vectors computed by PSB and by the hold-out grid search) to an AVHRR image acquired over Porto Alegre (Brazil) on January 15, 1999, at 3:42 P.M. local solar time [Fig. 2(d)]. We note that, in these maps, very low SST estimates were erroneously obtained in some pixels at the borders between sea and land or near cloud-covered areas due to the effect of mixed pixels.

D. Experiment III: Sensitivity to Initialization

PSB is iterative and is initialized with a predefined parameter vector. A further experiment was performed to assess the sensitivity of the method to this initialization. The initial parameter values were varied in the ranges $C^0 \in [0.1; 10]$, $\varepsilon^0 \in [0.001; 0.1]$, $\sigma^0 \in [0.1; 1]$, and a grid of 250 points was defined according to these ranges. For each of the three data sets, PSB was initialized to each node in the grid and was run until convergence. The MAE and rmse values, computed on the test sets and averaged over the 250 runs, are shown in Table IV, together with the related standard deviations.

When varying the initial parameter values in the aforementioned ranges, PSB converged, as expected, to different local minima of the span bound. However, a remarkable stability of the resulting test-set errors was noted. For all three data sets, the standard deviations of MAE and rmse were much smaller than the corresponding average values, which, in particular, were very close to the values reported in Tables I–III for the specific case $C^0 = 1$, $\varepsilon^0 = 0.01$, and $\sigma^0 = 0.5$. This is consistent with the behaviors of the span bound and of the test-set errors discussed in Section III-A, which presented large valleylike shapes including many local minima corresponding to different parameter configurations and to almost equivalent regression performances.

Fig. 3 shows the histograms of the computation times taken by PSB (applied to the three data sets) in the 250 runs. The average times were 29 min 55 s, 1 min 17 s, and 3 min 38 s in the cases of "LST-AVHRR-Mesonet," "LST-MSG-ARSIA," and "SST," respectively. Such times were shorter than the corresponding ones required by CV or hold-out grid searches (see Tables I–III). However, the choice of the initial parameter vector affected the time required by PSB to converge to a local minimum for each run. When considering the two LST data sets, the computation times of all PSB runs were (much) shorter than those of the grid searches. On the other hand, in the case of "SST," PSB took a longer time than CV in 57 out of 250 runs. In fact, thanks to the small training-sample size of this data set, CV required a short time. Such results further confirm that PSB, as compared with grid searches, yields a relevant reduction in the computational burden when many training samples are available, whereas this advantage may not be significant in small-sample-size cases, in which the execution time of grid searches is already noncritical.

E. Experiment IV: Bias-Variance Analysis

A specific experiment was carried out in order to investigate possible overfitting issues related to PSB. We recall that, given the regression problem consisting in the estimation of a random variable y as a function of an n-dimensional feature vector x, assuming a set $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ of i.i.d. training samples to be available, and denoting $\hat{f}_{\mathcal{D}}(\cdot)$ as the functional approximator obtained by training with $\mathcal{D}$ a supervised

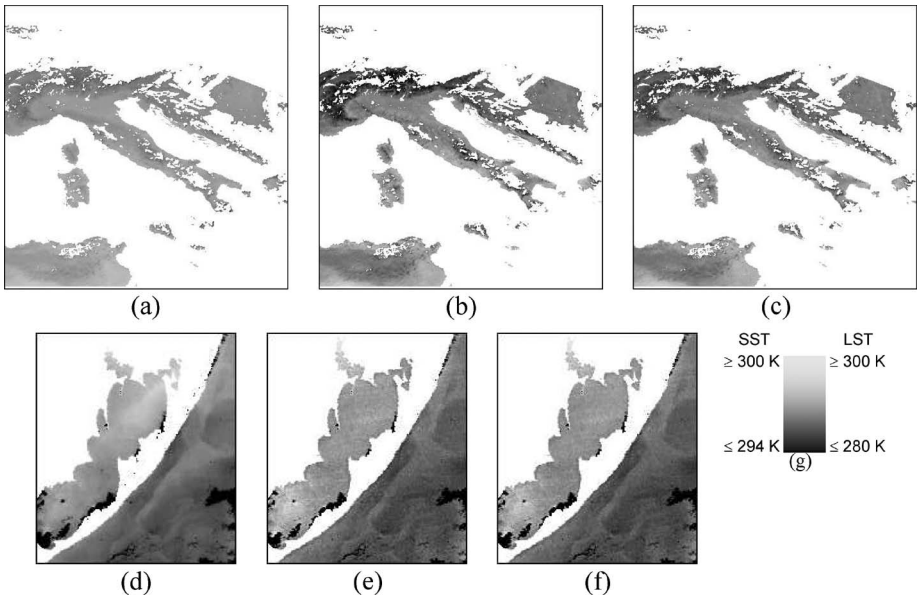

Fig. 2. Experiment II: (a) MSG image acquired over Italy on September 15, 1999, at 7:00 A.M. local time (the 10.8-μm brightness temperature channel is shown). (b) Map of LST estimates generated by SVM regression with parameters optimized by PSB. (c) Map of LST estimates generated by SVM regression with parameters optimized by the hold-out grid search. (d) AVHRR image acquired over Porto Alegre (Brazil) on January 15, 1999, at 3:42 P.M. local solar time (AVHRR channel 4 is shown). (e) Map of LST estimates generated by SVM regression with parameters optimized by PSB. (f) Map of LST estimates generated by SVM regression with parameters optimized by the hold-out grid search. (g) Color legend for the temperature maps. Cloud-covered areas are masked in white in all the images, together with sea areas in (a)–(c) and with land areas in (d)–(f).

TABLE IV
EXPERIMENT III: MEANS AND STANDARD DEVIATIONS (STDEV) OF THE MAE AND RMSE VALUES OBTAINED BY PSB ON THE TEST SETS OF THE THREE CONSIDERED DATA SETS, WHEN VARYING THE INITIAL PARAMETER VECTOR IN A PREDEFINED GRID OF 250 POINTS

		"LST-AVHRR-Mesonet"	"LST-MSG-ARSIA"	"SST"
MAE [K]	mean	1.10	1.53	0.57
	stdev	0.05	0.09	0.09
RMSE [K]	mean	1.43	1.97	0.80
	stdev	0.07	0.14	0.16

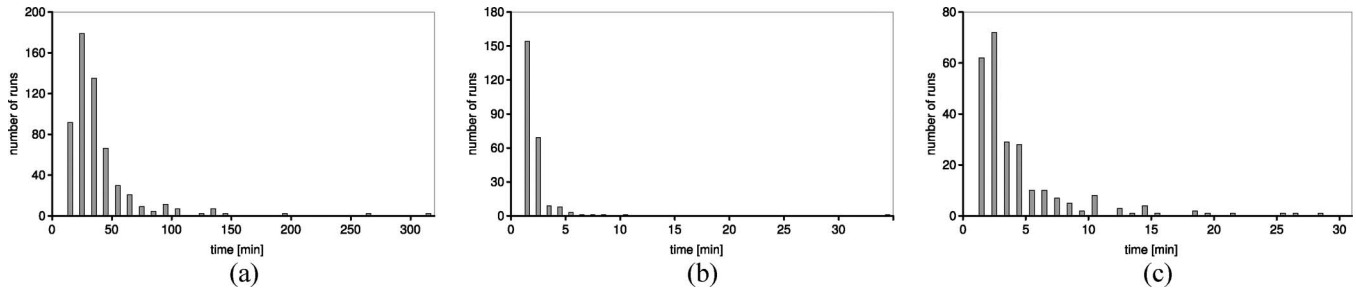

Fig. 3. Experiment III: Histograms of the computation times required by PSB to reach convergence, when varying the initial parameter vector in a predefined grid of 250 points. (a) "LST-AVHRR-Mesonet." (b) "LST-MSG-ARSIA." (c) "SST."

regression technique (e.g., an SVM or a neural network), the mse on an unknown test sample x can be decomposed as the sum of three contributions [51]

$$\begin{aligned}\mathrm{mse}(x) &= E\left\{\left[y-\hat{f}_{\mathcal{D}}(x)\right]^2|x\right\}\\ &= \mathrm{Var}\{y|x\}+\left[E\left\{\hat{f}_{\mathcal{D}}(x)|x\right\}-E\{y|x\}\right]^2\\ &\quad+\mathrm{Var}\left\{\hat{f}_{\mathcal{D}}(x)|x\right\}. \end{aligned} \tag{6}$$

The first term does not depend on the adopted estimator and is related to the intrinsic data variability (e.g., due to noise) of the target variable y. The second contribution represents a measure of the average bias of the adopted estimator $\hat{f}_{\mathcal{D}}(x)$ with respect to the conditional mean $E\{y|x\}$, which is known to be the optimal estimator in the mse sense [52] (and which is usually unknown). The third contribution is a measure of the variance of the estimator $\hat{f}_{\mathcal{D}}(x)$ as a function of the (random) training samples in $\mathcal{D}$. A tradeoff (namely, the so-called "bias-variance dilemma") often holds between the bias and variance terms [51]. For instance, sharply reducing the bias may cause overfitting phenomenon, which is likely to yield an increase in the variance.

For each of the three considered data sets, in order to perform a "bias-variance" analysis of PSB, 100 subsets were randomly

TABLE V
EXPERIMENT IV: SQUARE ROOTS OF THE INTEGRATED BIAS, VARIANCE, AND MSE COMPUTED ON THE TEST SET OF EACH OF THE CONSIDERED DATA SETS, ACCORDING TO THE LST/SST ESTIMATORS OBTAINED BY RUNNING PSB OVER 100 RANDOMLY SELECTED TRAINING SUBSETS

	"LST-AVHRR-Mesonet"	"LST-MSG-ARSIA"	"SST"
Root integrated bias [K]	1.56	1.92	0.72
Root integrated variance [K]	0.66	0.76	0.55
Root integrated MSE [K]	1.70	2.07	0.91

sampled from the available training set. According to the available training-sample sizes (Tables I–III), each subset was made of about 100 samples for "LST-AVHRR-Mesonet" and of about 50 samples for "LST-MSG-ARSIA" and "SST." PSB was initialized with $C^0 = 1$, $\varepsilon^0 = 0.01$, and $\sigma^0 = 0.5$ and was run until convergence over each training subset, thus yielding a distinct trained SVM estimator. Then, the "integrated bias," "integrated variance," and "integrated mse" were computed on the test set (see Table V), which represent estimates of the average bias, variance, and mse values. Details about such estimates can be found in [51]. Here, we only remark that they rely on the assumption that the mse contribution related to intrinsic data variability in (6) is negligible when compared to the other two terms. The resulting "integrated mse" is equal to the average of the 100 mse values obtained on the test set by the functional approximators trained on the 100 subsets.

We first note that the root integrated mse values in Table V were slightly higher than the rmse values in Tables I–III. This is probably because of the much smaller size of the training subsets involved in this experiment as compared to the original training-sample sizes. For all three data sets and, in particular, for the two LST ones, the resulting integrated biases were significantly larger than the corresponding integrated variances. The integrated bias was equal to 85%, 86%, and 63% of the corresponding integrated mses for "LST-AVHRR-Mesonet," "LST-MSG-ARSIA," and "SST," respectively. This suggests that the test-set errors can be mainly ascribed to the overall difference between the functional approximator generated by PSB and the (unknown) conditional mean of the target LST/SST given the feature values, and it is not because of large fluctuations in the approximator as the training samples are varied. This suggests a rather small sensitivity of PSB to the choice of specific sets of training samples, thus limiting the related risks of overfitting.

IV. CONCLUSION

In this paper, a novel automatic parameter optimization method has been proposed for SVM regression and validated in the context of LST and SST estimation from satellite infrared data. The method integrates the span-bound functional, developed in [35] as an upper bound on the LOO regression error, with Powell's numerical minimization algorithm in order to identify a quasi-optimal parameter configuration as compared with the criterion of the minimum estimation error. Experiments on three data sets, consisting of AVHRR and MSG satellite images with synchronous *in situ* measurements of LST and SST, pointed out strongly correlated behaviors of the aforesaid functional and of the MAE and rmse regression errors on test sets, thus confirming the appropriateness of the choice of this theoretical upper bound.

Experimental results on both SST and LST data sets generated by the proposed PSB method turned out to be very similar to those obtained by a classical grid-search approach based on CV on the sets of available training samples. Slightly worse results were obtained as compared with the ones provided by a further grid search for the parameter configuration yielding the smallest MAE on the test set, assuming the true *in situ* temperature values to be known also on test samples. However, the differences between the MAEs and rmses of this "ideal" benchmark case and the ones of the proposed technique were very small for all the considered data sets.

On the other hand, the computation time of PSB, when applied to the two LST data sets, turned out to be much shorter than the execution times of grid searches. In the case of SST, the difference between the times taken by the two approaches was lower due to the smaller number of training samples. This demonstrates the capabilities of PSB to jointly exhibit accuracies very close to those achieved by grid searches and to require a much lower computational burden. Note that the training sets employed here for experiments were made of a few hundred samples. Even in such cases, the grid searches required several hours to complete the training process, which makes them critical from a computational viewpoint, when applied with much larger training sets (the training time of an SVM is known to increase at least quadratically with the training-sample size [16]). In addition, a grid search for the lowest CV or hold-out error preliminarily requires the user/operator to manually predefine a search range and a quantization step on each parameter. On the other hand, the proposed method can be suggested as a feasible tool for fully automatic and much faster training with essentially equivalent accuracies and without the need for user interaction. This is important for an operational use of the SVM-based approach to LST/SST estimation in order to speed up the update of the functional approximator which is regularly used to generate maps of surface temperature estimates. From this perspective, the integration of the proposed approach with incremental learning procedures (e.g., [53]) is a future extension worth investigating.

The proposed method is iterative and is initialized with a starting parameter configuration. However, the choice of this initial configuration turns out not to be critical. It influences the specific local minimum to which Powell's algorithm converges, but without significantly affecting the test-set regression errors. The initial parameter vector has a stronger impact on the time taken by Powell's procedure to reach convergence. Regardless, in the cases of both LST data sets, this time was always significantly shorter than the one required by grid searches. In the SST experiments, the time of CV and the average time of PSB were comparable because of the aforementioned smaller number of training samples in this data set. An interesting possible extension of the proposed technique might lie in the

application of global minimization methods (e.g., genetic or evolutionary algorithms [54]) to the span-bound functional. This may prevent convergence to local minima, even though possibly increasing the computation time and requiring further parameters of the minimization procedure to be tuned. It might also be worth integrating the proposed LST estimation method with soil/atmosphere interface energy flux models (e.g., for flood prevention and monitoring) [55], [56].

A further analysis of the behavior of PSB from the viewpoint of the "bias-variance dilemma" was also performed by iteratively running PSB on randomly selected training subsets and then by estimating the integrated biases and variances of the resulting estimators. The experimental results suggested that the integrated bias contributions were dominant with respect to the integrated variances for the three data sets, which suggests that PSB has a limited sensitivity to the specific choice of the training samples (provided that they are drawn from the same probability distribution).

We note that the presented LST experiments referred to AVHRR and MSG images acquired between 7:00 and 9:00 A.M. (local times). Further experiments (not presented here for the sake of brevity) on MSG images acquired between 9:00 A.M. and 6:00 P.M. (local times) further confirm the similarity of the regression accuracies of the proposed technique to those of grid searches and the large differences in the related computation times. Furthermore, the choices of the features used in the regression experiments were suggested by traditional physically based split-window approaches [47], [48], [50]. As further developments of this paper, it would be interesting to validate the proposed method by considering, also in the LST case, the satellite zenith angle among the input features (thanks to its relationship with atmospheric absorption [13]) and, more generally, by automating the feature-extraction stage through the use of supervised (e.g., kernel-based [57]) algorithms to compute sets of features that optimize regression accuracy.

It is worth noting that, even though the proposed technique has been experimentally tested in the context of surface temperature estimation, it is not application specific. The span bound is a general bound on the LOO error of support vector regression. From a methodological viewpoint, this suggests that PSB is a general locally optimum parameter optimization algorithm for this family of regression techniques. On the other hand, a further validation on different regression problems (e.g., estimation of other bio-/geophysical parameters of the Earth's surface) would be required to appreciate the effectiveness of the method, particularly compared to grid searches, in other applicative contexts as well. This would be an important development of this paper. It would also be interesting to extend the method to SVM-based regression with quadratic penalty on the slack variables by combining the related span-bound formulation [35] with Powell's algorithm.

ACKNOWLEDGMENT

The authors would like to thank CIMA for providing the MSG images used for the experiments; Oklahoma Mesonet and NOAA for allowing free Web access to *in situ* temperature measurements in Oklahoma (USA) and to AVHRR data, respectively; ARSIA for acquiring temperature measurements in Tuscany (Italy), Prof. F. Castelli from the University of Florence (Italy) for providing such measurements; the AVHRR Pathfinder Oceans group of the University of Miami (USA) for allowing free Web access to the AVHRR Pathfinder Matchup Database of SST measurements; Dr. C. C. Chang and Prof. C. J. Lin from the University of Taipei (Taiwan) for freeware providing the LIBSVM software package; Dr. M. Zortea for his assistance in data preparation and preprocessing; and A. Agneessens for his help with the experiments.

REFERENCES

[1] P. Dash, F.-M. Gottsche, F.-S. Olesen, and H. Fischer, "Land surface temperature and emissivity estimation from passive sensor data: Theory and practice—Current trends," *Int. J. Remote Sens.*, vol. 23, no. 13, pp. 2563–2594, Jul. 2002.

[2] S. L. Haines, G. J. Jedlovec, and S. M. Lazarus, "A MODIS sea surface temperature composite for regional applications," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 9, pp. 2919–2927, Sep. 2007.

[3] K. Mao, J. Shi, H. Tang, Z.-L. Li, X. Wang, and K.-S. Chen, "A neural network technique for separating land surface emissivity and temperature from ASTER imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 1, pp. 200–208, Jan. 2008.

[4] L. F. Peres and C. C. DaCamara, "Improving two-temperature method retrievals based on a nonlinear optimization approach," *IEEE Geosci. Remote Sens. Lett.*, vol. 3, no. 2, pp. 232–236, Apr. 2006.

[5] J. C. Jimenez-Munoz and J. A. Sobrino, "Feasibility of retrieving land-surface temperature from ASTER TIR bands using two-channel algorithms: A case study of agricultural areas," *IEEE Geosci. Remote Sens. Lett.*, vol. 4, no. 1, pp. 60–64, Jan. 2007.

[6] I. J. Barton, "Satellite-derived sea surface temperatures—A comparison between operational, theoretical, and experimental algorithms," *J. Appl. Meteorol.*, vol. 31, no. 5, pp. 432–442, May 1992.

[7] P.-K. Chan and B.-C. Gao, "A comparison of MODIS, NCEP, and TMI sea surface temperature datasets," *IEEE Geosci. Remote Sens. Lett.*, vol. 2, no. 3, pp. 270–274, Jul. 2005.

[8] L. F. Peres and C. C. DaCamara, "Inverse problems theory and application: Analysis of the two-temperature method for land-surface temperature and emissivity estimation," *IEEE Geosci. Remote Sens. Lett.*, vol. 1, no. 3, pp. 206–210, Jul. 2004.

[9] L. F. Peres and C. C. DaCamara, "Emissivity maps to retrieve land-surface temperature from MSG/SEVIRI," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 8, pp. 1834–1844, Aug. 2005.

[10] C. T. Pinheiro, J. L. Privette, R. Mahoney, and C. J. Tucker, "Directional effects in a daily AVHRR land surface temperature dataset over Africa," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 9, pp. 1941–1954, Sep. 2004.

[11] A. P. Rodger, L. K. Balick, and W. B. Clodius, "The performance of the multispectral thermal imager (MTI) surface temperature retrieval algorithm at three sites," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 658–665, Mar. 2005.

[12] D. Sun and R. T. Pinker, "Case study of soil moisture effect on land surface temperature retrieval," *IEEE Geosci. Remote Sens. Lett.*, vol. 1, no. 2, pp. 127–130, Apr. 2004.

[13] Y. Yu, J. L. Privette, and A. C. Pinheiro, "Evaluation of split-window land surface temperature algorithms for generating climate data records," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 1, pp. 179–192, Jan. 2008.

[14] C. J. Merchant and P. L. Borgne, "Retrieval of sea surface temperature from space, based on modeling of infrared radiative transfer: Capabilities and limitations," *J. Atmos. Ocean. Technol.*, vol. 21, no. 11, pp. 1734–1746, Nov. 2004.

[15] M. Zortea, M. De Martino, G. Moser, and S. B. Serpico, "Land surface temperature estimation from infrared satellite data using support vector machines," in *Proc. IGARSS*, Denver, CO, Jul. 31,–Aug. 4, 2006, pp. 2109–2112.

[16] V. N. Vapnik, *Statistical Learning Theory*. New York: Wiley-Interscience, 1998.

[17] Y. Bazi and F. Melgani, "Semisupervised PSO-SVM regression for biophysical parameter estimation," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 6, pp. 1887–1895, Jun. 2007.

[18] L. Bruzzone and F. Melgani, "Robust multiple estimator systems for the analysis of biophysical parameters from remotely sensed data," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 1, pp. 159–174, Jan. 2005.
[19] G. Camps-Valls, L. Bruzzone, J. L. Rojo-Alvarez, and F. Melgani, "Robust support vector regression for biophysical variable estimation from remotely sensed images," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 3, pp. 339–343, Jul. 2006.
[20] G. Camps-Valls, L. Gomez-Chova, J. Munoz-Mari, J. Vila-Frances, J. Amoros-Lopez, and J. Calpe-Maravilla, "Retrieval of oceanic chlorophyll concentration with relevance vector machines," *Remote Sens. Environ.*, vol. 105, no. 1, pp. 23–33, Nov. 2006.
[21] F. Yang, M. A. White, A. R. Michaelis, K. Ichii, H. Hashimoto, P. Votava, A.-X. Zhu, and R. R. Nemani, "Prediction of continental-scale evapotranspiration by combining MODIS and AmeriFlux data through support vector machine," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 11, pp. 3452–3461, Nov. 2006.
[22] S. S. Durbha, R. L. King, and N. H. Younan, "Support vector machines regression for retrieval of leaf area index from multiangle imaging spectroradiometer," *Remote Sens. Environ.*, vol. 107, no. 1/2, pp. 348–361, Mar. 2007.
[23] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee, "Choosing multiple parameters for support vector machines," *Mach. Learn.*, vol. 46, no. 1–3, pp. 131–159, Jan. 2002.
[24] K.-M. Chung, W.-C. Kao, T. Sun, L.-L. Wang, and C.-J. Lin, "Radius margin bounds for support vector machines with the RBF kernel," *Neural Comput.*, vol. 15, no. 11, pp. 2643–2681, Nov. 2003.
[25] K. Duan, S. S. Keerthi, and A. N. Poo, "Evaluation of simple performance measures for tuning SVM hyperparameters," *Neurocomputing*, vol. 51, pp. 41–59, Apr. 2003.
[26] T. Joachims, "Estimating the generalization performance of an SVM efficiently," in *Proc. Int. Conf. Mach. Learn.*, 2000, pp. 431–438.
[27] G. Wahba, Y. Lin, and H. Zhang, "GACV for support vector machines," in *Advances in Large Margin Classifiers*, A. J. Smola, P. J. Bartlett, B. Scholkopf, and D. Schuurmans, Eds. Cambridge, MA: MIT Press, 1999.
[28] V. Vapnik and O. Chapelle, "Bounds on error expectation for support vector machine," in *Advances in Large Margin Classifiers*, A. J. Smola, P. J. Bartlett, B. Scholkopf, and D. Schuurmans, Eds. Cambridge, MA: MIT Press, 1999.
[29] C. Burges, *A Tutorial on Support Vector Machines for Pattern Recognition*. Boston, MA: Kluwer, 1998.
[30] P. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: Risk bounds and structural results," *J. Mach. Learn. Res.*, vol. 3, pp. 463–484, Mar. 2002.
[31] V. Cherkassky and Y. Ma, "Practical selection of SVM parameters and noise estimation for SVM regression," *Neural Netw.*, vol. 17, no. 1, pp. 113–126, Jan. 2003.
[32] J. T. Kwok and I. T. Tsang, "Linear dependency between ϵ and the input noise in ϵ-support vector regression," *IEEE Trans. Neural Netw.*, vol. 14, no. 3, pp. 544–553, May 2003.
[33] W. Chu, S. Keerthi, and C. J. Ong, "Bayesian support vector regression using a unified loss function," *IEEE Trans. Neural Netw.*, vol. 15, no. 1, pp. 29–44, Jan. 2004.
[34] J. B. Gao, S. R. Gunn, and C. J. Harris, "A probabilistic framework for SVM regression and error bar estimation," *Mach. Learn.*, vol. 46, no. 1–3, pp. 71–89, Jan. 2002.
[35] M.-W. Chang and C.-J. Lin, "Leave-one-out bounds for support vector regression model selection," *Neural Comput.*, vol. 17, no. 5, pp. 1188–1222, May 2005.
[36] J. Shawe-Taylor, C. K. I. Williams, N. Cristianini, and J. Kandola, "On the eigenspectrum of the Gram matrix and the generalization error of kernel-PCA," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2510–2522, Jul. 2005.
[37] W. H. Press, S. A. Teukolsky, W. T. Wetterling, and B. P. Flannery, *Numerical Recipes in C*. Cambridge, U.K.: Cambridge Univ. Press, 2002.
[38] M. Zortea, "Advanced pattern recognition techniques for environmental information extraction from remotely sensed data," Ph.D. dissertation, Univ. Genoa, Genoa, Italy, 2007.
[39] P. Mantero, G. Moser, and S. B. Serpico, "Partially supervised classification of remote sensing images through SVM-based probability density estimation," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 3, pp. 559–570, Mar. 2005.
[40] R.-E. Fan, P.-H. Chen, and C.-J. Lin, "Working set selection using second order information for training support vector machines," *J. Mach. Learn. Res.*, vol. 6, pp. 1889–1918, Dec. 2005.
[41] T. Joachims, "Making large-scale SVM learning practical," in *Advances in Kernel Methods—Support Vector Learning*, C. Burges and A. Smola, Eds. Cambridge, MA: MIT Press, 1999.
[42] M. J. D. Powell, "An efficient method for finding the minimum of a function of several variables without calculating derivatives," *Comput. J.*, vol. 7, no. 2, pp. 155–162, Feb. 1964.
[43] R. P. Brent, *Algorithms for minimization without derivatives*. Englewood Cliffs, NJ: Prentice-Hall, 1973.
[44] D. Pozo Vazquez, F. J. O. Reyes, and L. A. Arboledas, "A comparative study of algorithms for estimating land surface temperature from AVHRR data," *Remote Sens. Environ.*, vol. 62, no. 3, pp. 215–222, Dec. 1997.
[45] J. A. Sobrino, J. C. Jimenez-Munoz, M. El-Kharraz-Gomez, M. Romaguera, and G. Soria, "Single-channel and two-channel methods for land surface temperature retrieval from DAIS data and its application to the Barrax site," *Int. J. Remote Sens.*, vol. 25, no. 1, pp. 215–230, Jan. 2004.
[46] H. Ouaidrari, S.N. Goward, K.P. Czajkowski, J.A. Sobrino, and E. Vermote, "Land surface temperature estimation from AVHRR thermal infrared measurements: An assessment for the AVHRR Land Pathfinder II data set," *Remote Sens. Environ.*, vol. 81, no. 1, pp. 114–128, 2002.
[47] J. A. Sobrino and N. Raissouni, "Toward remote sensing methods for land cover dynamic monitoring: Application to Morocco," *Int. J. Remote Sens.*, vol. 21, no. 2, pp. 353–366, 2000.
[48] C. Ulivieri, M. M. Castronuovo, R. Francioni, and A. Cardillo, "A split window algorithm for estimating land surface temperature from satellites," *Adv. Space Res.*, vol. 14, no. 3, pp. 59–65, Mar. 1994.
[49] R. Evans and G. Podestá, "NOAA/NASA AVHRR oceans pathfinder sea surface temperature data set user's reference manual," NOAA/NASA, Washington, DC, 1998. Tech. Rep. [Online]. Available: http://www.rsmas.miami.edu/groups/ rrsl/pathfinder/
[50] X. Li, W. Pichel, E. Maturi, P. Clemente-Colon, and J. Sapper, "Deriving the operational nonlinear multichannel sea surface temperature algorithm coefficients for NOAA-15 AVHRR/3," *Int. J. Remote Sens.*, vol. 22, no. 4, pp. 699–704, 2001.
[51] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias/variance dilemma," *Neural Comput.*, vol. 4, no. 1, pp. 1–58, Jan. 1992.
[52] A. Papoulis and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes*. New York: McGraw-Hill, 2002.
[53] W. Wang, "An incremental learning strategy for support vector regression," *Neural Process. Lett.*, vol. 21, pp. 175–188, 2005.
[54] Z. Michalewicz, *Genetic algorithms + Data structures = Evolution programs*. New York: Springer-Verlag, 1996.
[55] G. Boni, F. Castelli, and D. Entekhabi, "Sampling strategies and assimilation of ground temperature for the estimation of surface energy balance components," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 1, pp. 165–172, Jan. 2001.
[56] F. Caparrini and F. Castelli, "Mapping of land–atmosphere heat fluxes and surface parameters with remote sensing data," *Boundary-Layer Meteorol.*, vol. 107, no. 3, pp. 605–633, Jun. 2003.
[57] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

Gabriele Moser (S'03–M'05) received the Laurea (M.S.) degree in telecommunications engineering (*summa cum laude*) and the Ph.D. degree in space sciences and engineering from the University of Genoa, Genoa, Italy, in 2001 and 2005, respectively.

Since 2001, he has been cooperating with the Signal Processing and Telecommunications Research Group (SP&T), Department of Biophysical and Electronic Engineering (DIBE), University of Genoa, in the field of remote sensing image analysis. From January to March 2004, he was a Visiting Student with the Institut National de Recherche en Informatique et en Automatique, Sophia Antipolis, France, working with the "Ariana" research group on the problem of SAR data modeling. He is currently a Research Fellow at DIBE. He is also with the Interuniversity Research Center in Environmental Monitoring (CIMA), Savona, Italy. His research activity is focused on image-processing and image-analysis methodologies for remote-sensing data interpretation. In particular, his current research interests include SAR data analysis, multitemporal image classification, hyperspectral image analysis, contextual classification, and geo-/biophysical parameter estimation.

Dr. Moser has been an Associate Editor for the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS since 2008. He has been a Reviewer for several international journals.

Sebastiano B. Serpico (M'87–SM'00) received the Laurea degree in electronic engineering and the Ph.D. degree in telecommunications from the University of Genoa, Genoa, Italy, in 1982 and 1989, respectively.

Since 1982, he has been cooperating with the Department of Biophysical and Electronic Engineering (DIBE), University of Genoa, in the field of image processing and recognition. He was an Assistant Professor from 1990 to 1998 and an Associate Professor of telecommunications from 1998 to 2004 with the Faculty of Engineering, University of Genoa, where he taught signal theory, pattern recognition, telecommunication systems, and electrical communications and is currently a Full Professor of telecommunications. From 1995 to 1998, he was the Head of the Signal Processing and Telecommunications Research Group (SP&T), DIBE, and is currently the Head of the SP&T Laboratory. He is the Chairman of the Institute of Advanced Studies in Information and Communication Technologies. He is also with the Interuniversity Research Center in Environmental Monitoring (CIMA), Savona, Italy. His current research interests include the application of pattern recognition (feature selection, classification, change detection, and data fusion) to remotely sensed images. He is the author or coauthor of more than 150 scientific publications, including journals and conference proceedings.

Dr. Serpico was the recipient of the Recognition of TGARS Best Reviewers from the IEEE Geoscience and Remote Sensing Society in 1998. He coedited a Special Issue of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING on the subject of the analysis of hyperspectral image data (July 2001) and a Special Issue on advances in techniques for analysis of remotely sensed data (March 2005). Since 2001, he has been an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. He is a member of the International Association for Pattern Recognition Society.