Modeling soil temperature based on Gaussian process regression in a semi-arid-climate, case study Ghardaia, Algeria

Redouane Mihoub · Nabil Chabour ·
Mawloud Guermoui

Received: 9 March 2016 / Accepted: 14 June 2016 / Published online: 1 July 2016
© Springer International Publishing Switzerland 2016

Abstract The renewable energy is the best energy potential to exploit, because they are economic, not pollutant and permanent. As a kind of renewable energy, geothermic which becomes more and more widely used in this field. In a geological setting regional on the effectiveness of the process solar thermal, offering a greater supply of geothermal energy, the study of Ghardaia's case are based on data of soil temperature and especially using local meteorological data, Accurate estimates of mean daily soil temperature (MDST) are needed. In this study, we will use the capability of Gaussian process regression (GPR) for modeling MDST using 3 years of measurement (2005–2008), in a semi-arid climate. It was found that GPR-model based on mean air temperature as input, give accurate results in term of mean absolute bias error, root mean square error, relative square error, and correlation coefficient. The obtained values

of these indicators are 0.0021, 0.5036, 0.0029 and 100 %, respectively, which shows that GPR is highly qualified for MDST estimation in semi-arid climate.

Keywords Soil · Temperature · Ghardaia · Meteorological · Solar radiations

1 Introduction

Algeria has a large area, the majority of which constitutes the Sahara. In the arid and semi-arid regions sunny places life the nomads. Algeria with its solar deposit very important to large opportunities for the development of the chain. Today, the country therefore has the duty to put in place an incentive policy in the framework of the operation and the popularization of these devices. The use of these technologies will open new prospects and will preserve the current reserves and to provide an alternative to oil and gas, from the point of view national income of the country and sources of energy.

The Sun emits electromagnetic radiation included in a the wave-length band ranging from 0.22 to 10 μm . The terrestrial atmosphere receives this radiation at an average power of 1.37 kW/m^2 to more or less 3 %. The amount of energy reaching the Earth's surface rarely exceeds 1200 W/m^2 . The rotation and the tilt of the Earth are also that the energy available in a given point varies depending on the latitude, the time and the

R. Mihoub (✉) · N. Chabour
Faculté des Sciences de la Terre, de la Géographie et de l'
Aménagement de Territoire, Université des Frères
Mentouri (Constantine - 1), BP.325 Route Ain el Bay,
25017 Constantine, Algeria
e-mail: redmihoub@yahoo.fr

N. Chabour
e-mail: nabilchabour@gmail.com

R. Mihoub · M. Guermoui
Unité de Recherche Appliquée en Energies
Renouvelables, URAER, Centre de Développement des
Energies Renouvelables, CDER, 47133 Ghardaïa, Algeria
e-mail: gur.mawloud@gmail.com

season. Clouds, fog, atmospheric particles and various other weather phenomena cause variations hourly and daily that increase or decrease the solar radiation.

The geological science is a part of renewable energy it consists in extracting the heat stored in the soil for the production of electricity, geothermal science in high temperature (Bastola and Peterson 2016), or the heating, geothermal science in low-temperature (Shamshirband et al. 2015). The temperature of the soil depends on the depth in which is measured, Such as the sun rays, the ambient temperature and the wind speed. The received solar radiation on the earth's surface depends upon the climatic conditions of a location and geological characteristics of the studied area (Wang and Bras 1999). An optimal use of solar energy (Yacef et al. 2014) needs an accurate knowledge of solar radiation at a particular geological location. Nevertheless, these data are not always dispensable particularly in isolated areas, in this respect several approaches have been developed in the literature for modeling (Yuan et al. 2008; Chan et al. 2013) and predicting soil temperature. Accurate measurement of soil temperature is a difficult task. Heat flux plates can be used to make direct measurements of soil temperature (Ni et al. 2014; Wang et al. 2011). At this stage, Gaussian process regression (GPR), relevance vector machine, and other methods. SVR, LSSVR, and GPR based soft sensors have attracted more attention recently because of their nonlinear modeling ability. However, the selection of suitable parameters for an SVR/LSSVR model is still difficult. Compared with SVR/LSSVR, the GPR model can optimize its parameters automatically. Additionally, GPR can simultaneously provide probabilistic information for its prediction; this is an appealing property in the process modeling area (Yi and Gao 2015).

However, the instruments of measure usually need to be placed at a certain depth in the soil normally a few centimeters below the surface according to avoid disturbances. Several single Gaussian process regression (GPR) models are first constructed for each steady-state grade (Yi et al. 2015) The objective of this work is to develop a simple method to model Soil Temperature based only on air temperature using Gaussian Process regression. The prediction can be achieved using the related steady-state GPR model if its reliability using this model is large is large enough (Yi et al. 2015). As the best of our knowledge this is

the first work that uses GPR for estimating the DST based only on DAR. The rest of this paper is organized as follows: Sect. 2 presents site location and data collection. In Sect. 3 we present the theory of GPR, model validation is presented in Sect. 4. Experimental results and discussion are presented in Sects. 5 and 6 concludes the and suggest a future work.

2 Site location and data collection

The experimental data used in this work (solar radiation, temperature, etc.), have been collected at the Applied Research Unit for Renewable Energies (URAER) situated in the south of Algeria (Fig. 1) far from the Ghardaïa city with latitude: (+32.370), longitude (+3.770), and altitude of (450 m) above the mean level.

The landscape is characterized by a vast expanse where rocky outcrops of bare rock a blackish brown color the values of soil (limestone) diffusivity are: $8.3910-7 \text{ m}^2/\text{s}$. This tray is masked by the strong river erosion early Quaternary who cut in its southern part of the flat-topped buttes and shaped valleys. The climate of Ghardaia region is semi-arid with a minimum and maximum air temperature ranging from 14 to 47 °C and from 2 to 37 °C during summer and winter months respectively. The daily global solar radiation (GSR) varies between a minimum of 607 Wh/m²/day to a maximum of 7574 Wh/m²/day and the annual-mean-daily GSR is about 5656 Wh/m²/day (Şenkal and Kuleli 2009). The data are recorded every 5 min with a high precision by a radiometric station installed at (URAER) (Fig. 2).

As mentioned above the prediction model uses the data collected between 2005 and 2008. The daily evolution of MDSR is shown in Fig. 3.

3 Theory of Gaussian process regression (GPR)

The theory of Gaussian process regression (GPR) has become increasingly a powerful statistical tool for data-driven modeling. GPR models are Bayesian non parametric approach that can be applied to solve classification and regression supervised (ML) problems. It has been applied to response surface modeling, system identification, calibration of spectroscopic analyzers (Vapnik and Vapnik 1998; Guermoui et al.

Fig. 1 Site location of Ghardaia city. **a** Algeria area, **b** Ghardaia area, **c** relief (3D) of Ghardaia area

Fig. 2 Instruments of measurement. Photo of the unit of Applied Research in Renewable Energies—URAER, Ghardaïa, **a** pyrheliometer, **b** pyranometer

2013) and ensemble learning. The main idea of GPR modeling is to place a prior directly on the space of functions. The combination of the prior and the data leads to the posterior distribution over functions. In this latter, we are focused on using the GPR approach for modeling (Sozen et al. 2004) the DGSR in the semi-arid area. Let us consider a regression x group containing d variables. In the machine-learning

Fig. 3 Daily evolution of normalized soil temperature

approach, the main objective is to learn the functional relationship between the inputs of (d-) dimensional ($\mathbf{x} \in \mathbb{R}^d$) and the output variable (y).

$$y = f(x) \quad (1)$$

where $(\mathbb{R})$ denotes the real space and f the unknown function. The unknown function f can be approximated by the following linear combination of basic function:

$$\hat{f}(x, w) = \sum_{j=1}^M w_j \phi_j(x) \quad (2)$$

$\{\phi_j(x)\}_{j=1}^M$, represent a set of basis function which can be linear or nonlinear and $w = [w_1, \dots, w_M]^T$ is the unknown vector for M basis function of (f) .

$$y = \sum_{j=1}^M w_j \phi_j(x) + \varepsilon, \quad (3)$$

In Eq. (3) ε represents the error term. In the general wide range of linear and nonlinear regression models uses a set of training data $(D = \{X, Y\}_{i=1}^N)$ Of (N) observation to estimate the unknown weights (w) and the basis function $(\phi_j(x))$ Can be seen as a transformation of the data from the original space in high dimensional space which is not the case in(GPR) models, as will be shown below. In their work (Suykens and Vandewalle 1999; Williams and Rasmussen 2006) mentioned that the basic block of GPR is a GP that assumes Gaussian priors for function values specified which is specified by its second order statistics:

$$f(x) \sim GP(m(x), k(x, x')) \quad (4)$$

where (x) , $k(x, x')$ represent the mean and the covariance function of f . By definition GP is a finite set of random variables with joint Gaussian distribution (Dong et al. 2005). Under GP, the prior distribution of (f) is Gaussian:

$$p(f|X, \theta) \sim \mathcal{N}(0, K) \quad (5)$$

The mean of f is assumed to be zero and the $N * N$ matrix K is a covariance matrix of f , with its hyper parameters denoted by θ .

If the error term ε in Eq. (5) is independent and identically Gaussian distributed, the likelihood function of the training target is also Gaussian:

$$p\{y|f, \sigma^2\} \sim \mathcal{N}(f, \sigma^2 I) \quad (6)$$

where σ^2 and I denote the variance of model error and identity matrix respectively. Then the posterior

distribution of f can be obtained by applying the Bayes' rule:

$$p(f|y, X, \theta, \sigma^2) = \frac{p(y|f, \sigma^2)p(f|X, \theta)}{p(y|X, \theta, \sigma^2)} \quad (7)$$

Note that the posterior distribution of f is also Gaussian, since both the prior and likelihood function is Gaussian. From (Suykens and Vandewalle 1999) the mean and covariance of the posterior distribution is given by:

$$\mu = K^T (K + \sigma^2 I)^{-1} \quad (8)$$

$$\Sigma = K - K^T (K + \sigma^2 I)^{-1} K \quad (9)$$

We note that the covariance function $K(\cdot, \cdot)$ is referred to us Kernel function in machine learning. In GPR literature come commonly used kernel functions include squared exponential or Gaussian kernel (Suykens and Vandewalle 1999).

$$k(x, x'|\theta) = \sigma_f^2 \exp\left(-\frac{r^2}{2l^2}\right), \theta = \{\alpha, l, \sigma_f^2\} \quad (10)$$

And the maternal family of covariance function is:

$$k(x, x'|\theta) = \sigma_f^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}r}{l}\right)^\nu k_\nu\left(\frac{\sqrt{2\nu}r}{l}\right), \theta\{\nu, l, \sigma_f^2\} \quad (11)$$

In Eqs. (10) and (11) the term $r = |x - x'|$ denote the Euclidean distance between two points and θ represent the hyper parameters associated with each covariance function. The variance noise (σ^2) is additional parameters that are determined during the training phase. The marginal probability distribution can be estimated by integration over the function f (Dong et al. 2005):

$$p(y|X) = \int p(y|f, \sigma^2)p(f|X, \theta) \quad (12)$$

The log marginal likelihood is obtained:

$$\log p(y|X) \propto -\frac{1}{2} y^T (K + \sigma^2 I)^{-1} y - \frac{1}{2} \log |K + \sigma^2 I| - \frac{N}{2} \log(2\pi) \quad (13)$$

Then the unknown parameters (θ, σ^2) can be estimated from the Eq. (13) using a gradient based algorithm. Since the posterior of f is determined through training data, we can evaluate the predictive

distribution of any test data (x_*) conditioned on training results:

$$p(f_*|x_*, y, X, \theta, \sigma^2) \quad (14)$$

From (Suykens and Vandewalle 1999) it can be shown that the predictive distribution Eq. (14) is Gaussian with mean m and variance (v^2) given by:

$$m(x_*) = \phi(x_*)^T \mu = K_*^T (K + \sigma^2 I)^{-1} y \quad (15)$$

$$v^2(x_*) = \phi(x_*)^T \Sigma \phi(x_*) = K_{**} - K_*^T (K + \sigma^2 I)^{-1} K_* \quad (16)$$

$K_* = [K(x_*, x_1), \dots, K(x_*, x_N)]^T$, $K_{**} = K(x_*, x_*)$, μ and Σ are the posterior mean and variance of f . The prior mean was assumed to zeros and the kernel function used in the present work is squared exponential.

4 Model validation

In this latter, the performance of GPR modeling of DGSR on a horizontal surface is evaluated by comparing the estimated values with these measured using different statistical indexes such as mean absolute bias error (MABE), root mean square error (RMSE), relative square error (RRMSE), determination coefficient (R^2) and correlation coefficient (r): The MABE, give the mean absolute value of bias error. Its expression is given Eq. (17) by:

$$MABE = \frac{1}{n} \sum_{i=1}^n |H' - H| \quad (17)$$

where (H') Is the estimated value and (H) is the measured value and ($i = 1, \dots, n$) number of observations.

The RMSE represents the difference between the predicted values and the measured values. In fact RMSE identifies the model's accuracy. It is calculated Eq. (18) by:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (H' - H)^2} \quad (18)$$

The RRMSE is calculated by dividing the RMSE to the average of measured data as:

$$RRMSE = \frac{\sqrt{\frac{1}{n} \sum_{i=1}^n (H' - H)^2}}{\frac{1}{N} \sum_{i=1}^N H} \times 10 \quad (19)$$

The performance of the model is defined by the RRMSE range as follows:

Excellent if: $RMSE < 10\%$
 Good if: $10\% < RMSE < 20\%$
 Fair if: $20\% < RMSE < 30\%$
 Poor if: $RMSE > 30\%$

The r indicate the strength of a linear relationship between the measured and predicted values:

$$r = \frac{\sum_{i=1}^n (H' - \bar{H}') \cdot (H - \bar{H})}{\sqrt{\sum_{i=1}^n (H' - \bar{H}')^2 \cdot \sum_{i=1}^n (H - \bar{H})^2}} \quad (20)$$

5 Experimental results

In this section, we will introduce the application of GPR for modeling MDSR using MDAT as input and MDST as output. Usually, measuring such physical quantities would include formulas that mathematically describe the relationships between the parameters inputs and the desired output. The experimental database used in the current study contain 1061 days of measurement. For the training of GPR model we are splitting the database into tow subset. The first subset contain 560 days for training and the second one 501 days for testing the model.

As shown in Fig. 4, we observe that GPR model based on air temperature as input give high precision and the predicted values of MDST are similar to the measured values.

An important observation from Fig. 5 is that using air temperature alone as input achieves high performance due to its high correlation with the soil temperature.

The obtained statistical indexes confirm also the performance of the proposed model. The values of these indexes (Table 1).

Now that we have all the information regarding the prior and the hierarchical priors, for a given new point(x), (Fig. 5) the value of the target variable can be predicted as The resulting expression.

6 Conclusion

In this work we present the applicability of Gaussian process regression for modeling soil temperature using

Fig. 4 Predicted and measured values of soil temperature

Fig. 5 Correlation coefficient (observed soil temperature and predicted soil temperature ($x = 1 \times x - 0.00034$))

Table 1 The obtained statistical indexes

Index	MAE	MABE	RMSE	Corr-Coeff
Results	0.5036	0.0021	0.0029	100 %

only air temperature as input. The obtained result is very satisfactory this due to the high correlation between the input and the output and the good precision of GPR for modeling the no linear relationship between the soil and air temperature compared with other recent models such as neural networks and support vector machine.

As a perspective to this work, we will use GPR for modeling soil temperature at different depth of soil using other available meteorological data.

References

- Bastola H, Peterson EW (2016) Heat tracing to examine seasonal groundwater flow beneath a low-gradient stream in rural central Illinois, USA. *Hydrogeol J* 24(1):181–194
- Chan et al (2013) Nonlinear system identification with selective recursive Gaussian process models. *Ind Eng Chem Res* 52(51):18276–18286
- Dong B et al (2005) Applying support vector machines to predict building energy consumption in tropical region. *Energy Build* 37(5):545–553
- Guermoui et al (2013) Heart sounds analysis using wavelets responses and support vector machines. In: 2013 8th international workshop on systems, signal processing and their applications (WoSSPA). IEEE, pp 233–238
- Liu Y, Gao Z (2015) Industrial melt index prediction with the ensemble anti-outlier just-in-time Gaussian process regression modeling method. *J Appl Polym Sci* 132(22). doi:10.1002/app.41958

- Liu Y, Chen T, Chen J (2015) Auto-switch Gaussian process regression-based probabilistic soft sensors for industrial multigrade processes with transitions. *Ind Eng Chem Res* 54(18):5037–5047
- Ni et al (2014) Non-linear calibration models for near infrared spectroscopy. *Anal Chim Acta* 813:1–14
- Şenkal O, Kuleli T (2009) Estimation of solar radiation over Turkey using artificial neural network and satellite data. *Appl Energy* 86(7):1222–1228
- Shamshirband et al (2015) Daily global solar radiation prediction from air temperatures using kernel extreme learning machine: a case study for Iran. *J Atmos Solar Terr Phys* 134:109–117
- Sozen A et al (2004) Estimation of solar potential in Turkey by artificial neural networks using meteorological and geographical data. *Energy Convers Manag* 45(18):3033–3052
- Suykens JAK, Vandewalle J (1999) Least squares support vector machine classifiers. *Neural Process Lett* 9(3):293–300
- Vapnik VN, Vapnik V (1998) *Statistical learning theory*, vol 1. Wiley, New York
- Wang J, Bras RL (1999) Ground heat flux estimated from surface soil temperature. *J Hydrol* 216(3):214–226
- Wang et al (2011) Raymond Bagging for robust non-linear multivariate calibration of spectroscopy. *Chemometr Intell Lab Syst* 105(1):1–6
- Williams CK, Rasmussen CE (2006) *Gaussian processes for machine learning*, vol 2. MIT Press, p 4
- Yacef et al (2014) New combined models for estimating daily global solar radiation from measured air temperature in semi-arid climates: application in Ghardaia, Algeria. *Energy Convers Manag* 79:606–615
- Yuan et al (2008) Reliable multi-objective optimization of high-speed WEDM process based on Gaussian process regression. *Int J Mach Tools Manuf* 48(1):47–60