5th International Conference of the International Commission of Agricultural and Biosystems Engineering (CIGR)

Hosted by the Canadian Society for Bioengineering (CSBE/SCGAB)
Québec City, Canada - June 14-18, 2020

# A NEW COUPLING VECTOR MACHINE APPROACH AND WAVELET TRANSFORM MODEL FOR DAILY SOIL TEMPERATURE PREDICTION AT DIFFERENT DEPTHS

MOHAMMAD ZEYNODDIN[1], HOSSEIN BONAKDARI[2]

[1]Department of Soils and Agri-Food Engineering, Laval University, Québec, Canada < Mohammad.zeynoddin.1@ulaval.ca>
[2]Department of Soils and Agri-Food Engineering, Laval University, Québec, Canada <hossein.bonakdari@fsaa.ulaval.ca >

**ABSTRACT** Soil temperature plays a critical role in the fields of hydrology, environmental science, ecology, soil science, meteorology, geotechnics and agronomy. In the present study the main purpose is to evaluate the performances of a novelty structured hybrid support vector machine model, in prediction of daily soil temperatures at six different depths from 5 to 100 cm. Wavelet analysis is also applied in order to pre-process the time series of meteorological data and obtain more accurate results. The modeling is carried out using the widely available input variables of maximum air temperature ($T_{max}$), minimum air temperature ($T_{min}$), evaporation from pan (EP) and sunshine duration (SD). The results are compared with Multi-Layer Perceptron Artificial Neural Network (MLP-ANN). All the criteria indicated that the results obtained by the SVM-WAVE are more precise than those of the MLP-ANN. The SVM-WAVE by average indices of coefficient of determination $R^2$ 0.953, Root mean squared error RMSE 1.928, Mean absolute error MAE 1.452, Mean absolute percentage error MAPE 9.970%, Average performance error APE 6.318% and Nash-Sutcliffe efficiency NS 94.865 outperformed the MLP-ANN by $R^2$ 0.947, RMSE 2.407, MAE 1.858, MAPE 11.913%, APE 8.123% and NS 93.022. The proposed method shows promising results for obtaining precise ST models with comprehensible conventional methods.

**Keywords:** Artificial Neural Network, Prediction, Soil temperature, Support vector machine, Wavelet transform, Meteorological variables.

**INTRODUCTION** Soil temperature (ST) plays a critical role in the fields of hydrology, environmental science, ecology, soil science, meteorology, geotechnics and agronomy (Jackson et al., 2008). Soil temperature as a fundamental meteorological variable, is related to various environmental factors including meteorological conditions (such as air temperature, evaporation rate and global/surface solar radiation), topographical variables (such as elevation, slope and aspect), soil physical variables (such as texture, albedo of surface, soil moisture and water content), and the other surface attributes (Zeynoddin et al. 2019). ST can influence rates of hydrological process interactions between land surface and atmosphere (Hu and Feng, 2003) and also different ecosystems from desert to forest (Jebamalar et al., 2012). Prediction of soil surface temperature has

a significant impact on the numerical hydrological and meteorological models (Gao et al., 2008).

Concerning impact of ST in various fields, developing forecasting methods for ST is essential, whether for surface soil or for soil at different depths. By reviewing the literatures of the subject, many methods have concentrated on development of analytical, mathematical, conceptual, empirical, experimental and numerical models (Yılmaz et al., 2009). The A methods as data-driven models are powerful computational techniques that as Haykin (2009) stated and they are primarily used for pattern recognition, classification and time series prediction. The Artificial Neural Networks (ANNs), known as black-box tools, are able to find the patterns even in intricate systems and they are easily applied in simulation of dynamic nonlinear systems (Rezaeian-Zadeh et al., 2012). The Support Vector Machines (SVMs) also have wide range of application in hydrological and environmental estimating such as solar radiation, sediment, evapotranspiration and flow (Chen et al., 2011). Compared to conventional learning approaches like neural networks, the SVMs have better generalization abilities for not-sampled data. Wavelet transforms (WT) is more efficient than the Fourier transforms for pre-processing of non-stationary time series (Kaheil et al. 2008). Fourier transforms is only able to obtain the frequency information of an original signal, whereas wavelet transforms has feature obtaining information on the time, location and frequency of that signal altogether. In the present study the main purpose is to evaluate the accuracies and performances of structured hybrid SVMs model, namely SVM-WAVE and a MLP-ANN model in prediction of daily soil temperatures at six different depths. In this project, Wavelet analysis is applied in order to decompose the ST time series and use decomposed components as the inputs for the SVMs model. For this research, the 4 daily meteorological variables of Zahak station in southeast of Iran are used to develop models. Eventually, these two developed methods are compared together and their efficiencies are specified through various statistical criteria.

## MATERIALS AND METHODS DESCRIPTION

**Support vector machine (SVM)** The SVM method (Vapnik, 2000) is one of the most common neural networks that is used successfully in various fields of engineering for forecasting, classification, pattern recognition, and regression analyses (Lee and Verri, 2003; Lu and Wang, 2005; Asefa et al., 2006; Ji and Sun, 2013; Sun, 2013). From the Vapnik theory, the SVM functions are presented in Eqs. (1-4). In this method, the data points are shown by x and output by f(x), where xi is the input parameter, di is the target value, and n is the size of di. The functions are estimated by using Eqs. (1 and 2).

$$f(x) = w\varphi(x) + b \tag{1}$$

$$R_{SVMs}(C) = \frac{1}{2}\|w\|^2 + C\frac{1}{n}\sum_{i=1}^{n} L(x_i, d_i) \tag{2}$$

where $\varphi(x)$ is the mapped high dimensional space from *x*, *w* is the normal vector and *b* is the scalar vector. The second term of Eq. (2) stands for the risk or error. The *b* and *w* vectors are measured by minimizing the regularized risk equation as follows:

$$\text{Minimize} \quad R_{SVMs}\left(w,\xi^{(*)}\right)=\frac{1}{2}\|w\|^2+C\sum_{\mathrm{i}=1}^{\mathrm{n}}(\xi_i+\xi_i^*) \tag{3}$$

$$\text{Subject to} \quad \begin{cases} d_i-w\varphi(x_i)+b_i\le\varepsilon+\xi_i \\ w\varphi(x_i)+b_i-d_i\le\varepsilon+\xi_i^* \\ \xi_i,\xi_i^*\ge 0, i=1,\ldots,l \end{cases} \tag{4}$$

where $\xi_i$ and $\xi_i^*$ are slack variables, $\frac{1}{2}\|w\|^2$ is the regularization term, $\varepsilon$ is the loss function, *l* is the quantity of the training data's factor, and *C* is the error penalty to show the trade-off between the regularization term and the empirical error. To solve Eq. (1), the optimality constraints and Lagrange multiplier are used according to the following equation:

$$f\left(x,a_i a_i^*\right)=\sum_{\mathrm{i}=1}^{\mathrm{n}}(a_i-a_i^*)K(x,x_i)+b \tag{5}$$

where *K* is the kernel function that is calculated by using two inner vectors, $x_i$ and $x_j$, in future spaces of $\varphi(x_i)$ and $\varphi(x_j)$ by using the equation $K(x,x_i)=\varphi(x_i)\varphi(x_j)$.

The kernel function uses nonlinear mapping in order to have the ability to operate in higher dimensional space. In the SVM procedure, four different kernel functions of sigmoid, polynomial, Radial Basis Function (RBF), and linear are commonly used. However, various studies have shown that the RBF kernel function has better performance when facing complex problems (Rajasekaran et al., 2008; Yang et al., 2009; Wu and Wang, 2009). The RBF kernel function is defined as follows:

$$K(x_i,x_j)=\exp\left(-\gamma\|x_i-x_j\|^2\right) \tag{6}$$

where $x_i$ and $x_j$ are the input space vector and the parameter *γ* is calculated by $\gamma=-\frac{1}{2\sigma^2}$ where *σ* is the standard deviation's Gaussian noise level. In the present study, the RBF kernel function is applied. Selecting the appropriate values of *C*, *γ* and *ε* directly affects the accuracy of the results. Therefore, in the present study, the trial and error method is used in order to choose the correct values for these parameters.

**Wavelet Transform (WT)** This mathematical expression is used to decompose time series into different components. WT aims to increase the study model's size in order to have

access to the information at different levels (Adamowski and Chan, 2011). By using WT on a time domain, the performance of time domain localization is increased because of the ability to extract information from non-periodic and transient signals (Jawerth and Sweldens, 1994).

There are several WT fundamental functions that should be designed based on the case that we are studying. One of the WT subsets that are used for time-scale signal processing is Continuous Wavelet Transform (CWT). In CWT, the integral of all signals over the entire time period is calculated.

$$W_x\left(a,b,\psi\right)=\frac{1}{\sqrt{a}}\int_{-\infty}^{+\infty}f\left(t\right)\psi^*\left(\frac{t-b}{a}\right)dt \tag{7}$$

where $f(t)$ is the signal, $\psi(t)$ is the mother wavelet function, $a$ is the scale parameter, $\psi^*(t)$ is the multifaceted conjugate of $\psi$, $b$ is the time shift, and $t$ is time. Using the following equations for calculating the $a$ and $b$ parameters and discretizing Eq. (6), the Discrete Wavelet Transform (DWT) is obtained.

$$a=a_0^m\ ,\ b=na_0^mb_0\ ,\ a_0>1\ ,\ b_0\in R \tag{8}$$

where $n$ and $m$ are integers. The mathematical relation of DWT is obtained by substituting $a$ and $b$ from Eq. (7) in Eq. (6) as follows:

$$W_x\left(m,n,\psi\right)=a_0^{-m/2}\int_{-\infty}^{+\infty}f\left(t\right)\psi^*\left(a_0^{-m}t-nb_0\right)dt \tag{9}$$

where $W_x\left(a,b,\psi\right)$ shows the characteristics of the original time series for $a$ frequency and $b$ time domain. Mostly, $b_0$ and $a_0$ are selected for 1 and 2 time steps, respectively. A small $a$ leads to low frequency resolution of wavelet transform and a high time domain resolution.

The SVM based WT model's flowchart is presented in Fig. 1. According to this figure, in the present analysis, five different wavelet components are used. After applying the wavelet transformation, the decomposed components are chosen as the inputs of the SVM algorithm. Finally, the SVM results are combined in order to obtain the final results of the SVM-WAVE method.

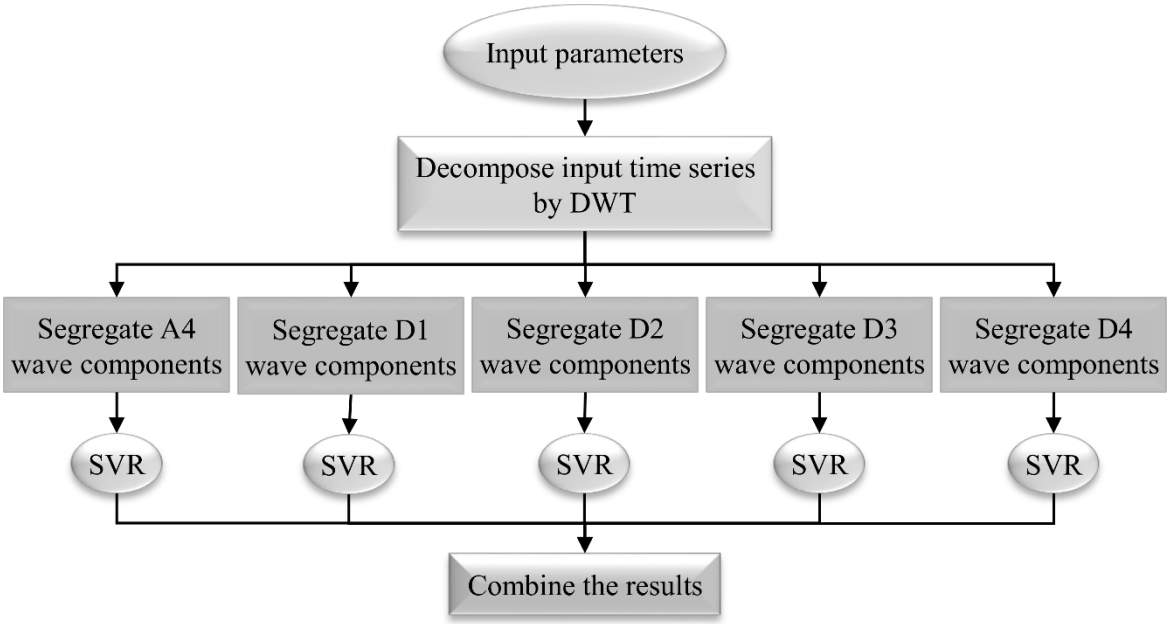


Figure 1. The SVM-WAVE flowchart

**DATA DESCRIPTION AND PROCESSING** In the present research, the meteorological variables were measured at Zahak synoptic meteorological station, which is located in the city of Zahak in Sistan and Baluchestan Province in the southeast of Iran. The zone map and the location of the station are shown in Fig. 2. The geographical coordinates for this station are 30°53'53" N latitude, 61°40'31" E longitude with an altitude of nearly 495 m above mean sea level (AMSL). The weather condition in the study zone is arid and desert. Also, the land surface is not covered by any specific types of vegetation, so there is a bare soil condition.


Figure 2. The zone map and the location of Zahak station in Sistan and Baluchestan Province in Iran

The daily measured meteorological data records including 702 days (i.e., 23 months) have been obtained from Zahak station. The easily obtained and mostly available meteorological variables consisting of maximum air temperature ($T_{max}$), minimum air temperature ($T_{min}$), pan evaporation (EP), sunshine duration (SD) and soil temperature (ST) at different depths of 5, 10, 20, 30, 50 and 100 cm beneath ground level, have contributed to the development of both SVM-WAVE and MLP-ANN models. The time series data of the daily measured STs at Zahak station corresponding to various depths are displayed in Fig. 3.

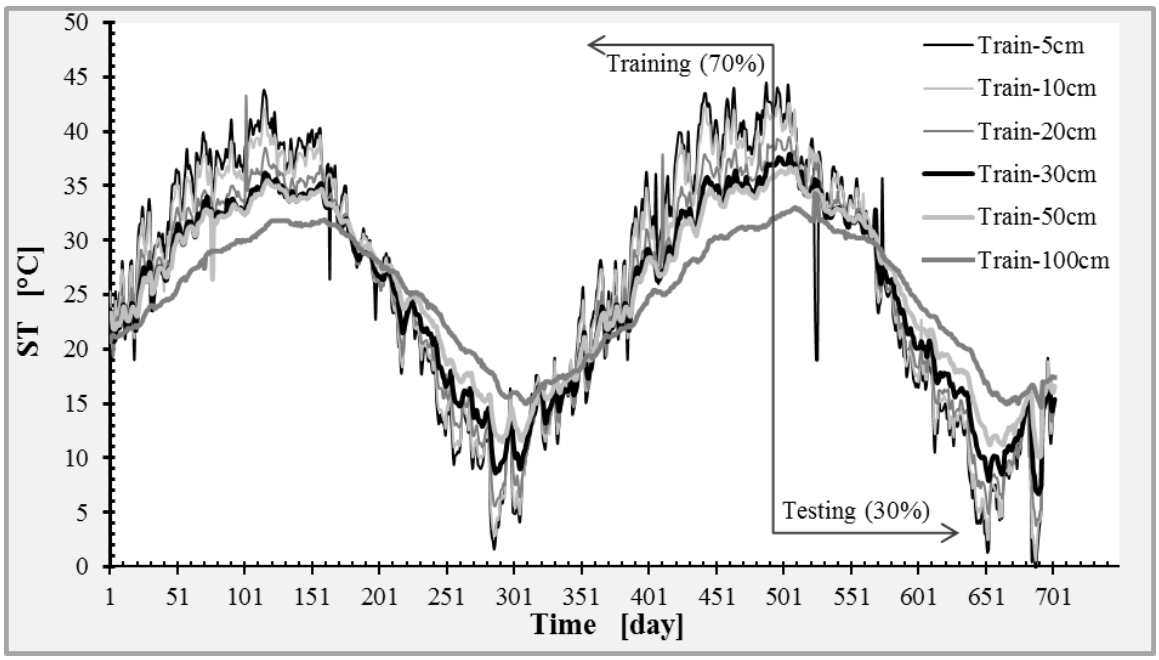

Figure 3. Time series data of the daily measured ST for Zahak station corresponding to various depths

Some of the most useful descriptive statistics for the utilized meteorological data sets in the study have been presented in Table 1. These important indices are determined separately for both the training and the testing data sets to demonstrate a better analysis. These indices consist of $O_{max}$ (maximum), $O_{min}$ (minimum), $O_{avg}$ (average), $S_d$ (sample standard deviation), $C_v$ (coefficient of variations) and $C_s$ (skewness coefficient) for each specific data set, which are listed in Table 1.

Table 1. The important statistical parameters of the utilized data sets

| Parameters | Data set | Tmax | Tmin | EP | SD | ST at depths (°C) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | (°C) | (°C) | (mm) | (hr) | 5cm | 10cm | 20cm | 30cm | 50cm | 100cm |
| Omax | Training | 47.00 | 33.40 | 30.00 | 13.20 | 44.50 | 42.50 | 43.30 | 37.00 | 35.70 | 31.90 |
| | Testing | 47.00 | 33.00 | 30.00 | 12.10 | 44.30 | 42.60 | 39.50 | 37.90 | 36.70 | 34.30 |
| Omin | Training | 4.60 | -6.20 | 0.00 | 0.00 | 1.60 | 2.90 | 5.60 | 8.60 | 11.50 | 15.00 |
| | Testing | -3.60 | -6.60 | 0.00 | 0.00 | -0.20 | 0.60 | 3.80 | 6.70 | 10.10 | 14.70 |
| Oavg | Training | 32.54 | 17.29 | 12.59 | 9.69 | 28.24 | 27.68 | 26.50 | 26.15 | 26.43 | 25.17 |
| | Testing | 27.61 | 12.79 | 9.16 | 8.69 | 21.57 | 22.01 | 22.00 | 22.77 | 23.89 | 24.56 |
| Sd | Training | 9.16 | 9.25 | 8.15 | 2.72 | 10.89 | 10.17 | 8.85 | 7.80 | 6.95 | 5.12 |
| | Testing | 12.18 | 10.56 | 8.18 | 2.92 | 12.54 | 12.18 | 10.89 | 9.78 | 8.71 | 6.45 |
| Cv | Training | 0.28 | 0.53 | 0.65 | 0.28 | 0.39 | 0.37 | 0.33 | 0.30 | 0.26 | 0.20 |
| | Testing | 0.44 | 0.83 | 0.89 | 0.34 | 0.58 | 0.55 | 0.50 | 0.43 | 0.36 | 0.26 |
| Cs | Training | -0.60 | -0.44 | 0.30 | -1.61 | -0.53 | -0.56 | -0.55 | -0.53 | -0.50 | -0.37 |
| | Testing | -0.20 | 0.21 | 0.82 | -1.74 | 0.18 | 0.12 | 0.08 | 0.04 | -0.01 | -0.15 |

**OPERATIONS AND RESULTS** In both the hybrid SVM-WAVE and the MLP-ANN methods, the inputs are exactly the four widely available and popular daily meteorological parameters: Tmin , Tmax , EP and SD. On the other hand, the outputs are the daily STs at six depths of 5, 10, 20, 30, 50 and 100 cm at time of t. So, it can be stated mathematically as:

$$ST^t = F\ (T^t_{\max},\ T^t_{\min},\ EP^t,\ SD^t) \tag{10}$$

where index t indicates that the models receive all four inputs relating to the particular time or day and consequently, the corresponding STs at the same time are provided. For the present methodologies, only a unique combination of available inputs, namely $\mathrm{T_{max}} + \mathrm{T_{min}} + \mathrm{EP} + \mathrm{SD}$, have been examined, since all four inputs are the most effective and commonly tried variables in the study literature (Bilgili, 2010; Ozturk et al., 2011; Napagoda and Tilakaratne, 2012) and approximately it has not been reported yet that a

prediction modeling of the daily ST using the hybrid SVM-WAVE and the MLP-ANN methods has been done. Main advantages of the proposed models is that they don't need the historical data inputs for prediction and hierarchy does not affect the models. In order to evaluate and compare the ability of both models six various statistical indicators are used, containing Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Nash-Sutcliffe efficiency (NS), Average Performance Error (APE) and coefficient of determination ($R^2$). The equations of the mentioned criteria have been presented in Table 2. Obviously, the criteria MAE, RMSE, MAPE and APE, expressing amount of the uncertainty in prediction, show better performances of the methods with low values, while for the criteria NS and $R^2$, the higher values present the better performances.

Table 2. The employed mathematical formulas of performance evaluation statistical criteria

| Crit. | Equation | Crit. | Equation |
|---|---|---|---|
| RMSE$^{oC}$ | $\sqrt{(1/N)\sum_{i=1}^{N}(O_i-P_i)^2}$ | NS% | $100(1-(\sum_{i=1}^{N}(O_i-P_i)^2/\sum_{i=1}^{N}(O_i-\overline{O})^2))$ |
| MAE $^{oC}$ | $(1/N)\sum_{i=1}^{N}\lvert O_i-P_i\rvert$ | APE% | $100(\sum_{i=1}^{N}\lvert O_i-P_i\rvert)/\sum_{i=1}^{N}O_i)$ |
| MAPE% | $(\frac{100}{N})\sum_{i=1}^{N}\lvert\frac{(O_i-P_i)}{O_i}\rvert$ | $R^2$ | $\frac{\sum_{i=1}^{N}(O_i-\overline{O})^2(P_i-\overline{P})^2}{(\sum_{i=1}^{N}(O_i-\overline{O})^2\sum_{i=1}^{N}(P_i-\overline{P})^2}$ |

In these formulas, $O_i$ , $P_i$ , $\overline{O}$ , $\overline{P}$ , and N represent the observed daily $ST_s$, the predicted daily $ST_s$, the mean observed daily $ST_s$, the mean predicted daily $ST_s$ and the number of days or data, respectively. Figure 4 presents the models vs. observed data for all depths. It can be concluded that both models can forecast $ST_s$ data with good accuracy. However, the hybrid SVM-WAVE produced more precise results than those estimated by the MLP-ANN. Furthermore, as the depth increase, the precision for both models decreased.

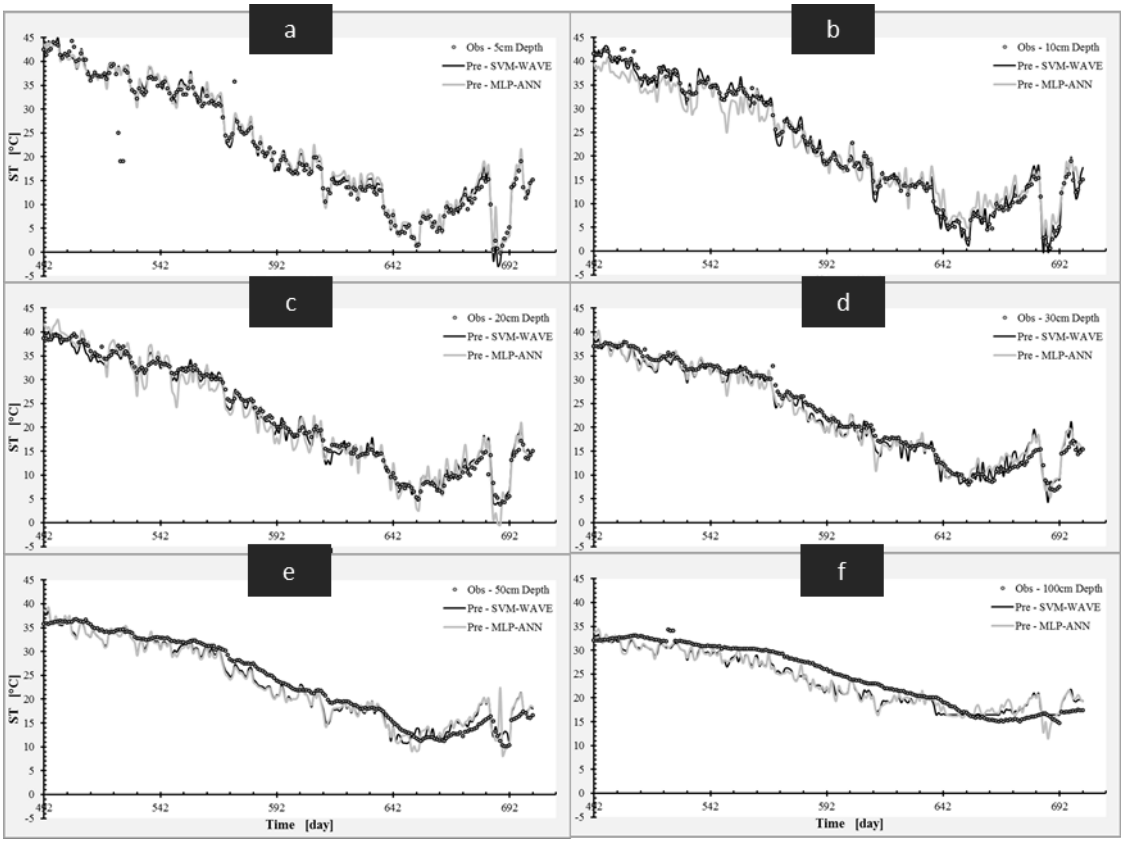


Figure 4. Comparison between obtained results of the SVM-WAVE and the MLP-ANN models in contrast to the observed daily ST based on the testing data, respectively for depths of (a) 5cm; (b) 10cm; (c) 20cm; (d) 30cm; (e) 50cm and (f) 100cm

In Table 3, the evaluation indices are provided for predictions of both AI models for test period. Regarding the introduced criteria, the predictions of hybrid SVM-WAVE are better and more accurate than the MLP-ANN except for depth 100 cm. Nonetheless, the hybrid SVM-WAVE has the lowest errors with respect to the criteria RMSE, MAE, MAPE and APE, and has the highest efficiencies with respect to the criterion NS in comparison to the MLP-ANN model. According to Table 3 testing results, the best ones relating to RMSE, MAE, MAPE, NS and APE criteria are 1.22°C, 0.98°C, 7.88%, 98.99% and 4.45% for the SVM-WAVE, and 1.99°C, 1.46°C, 9.08%, 95.85% and 6.77% for the MLP-ANN, at different depths respectively. These criteria show the superiorities of the hybrid SVM-WAVE at each of depths, except for depth 5 cm. The MAPE indicates better result for the MLP-ANN at this depth which can be concerning the uncertainty and complications in measurement of soil temperature at surface. consequently, it can be concluded that all the five evaluation indicators strongly confirm the better accuracies and performances of the hybrid SVM-WAVE compared to the most conventional MLP-ANN model. As mentioned earlier, increase in depth, results in precision decline.For example, concerning MAE index, accuracy of the hybrid SVM-WAVE declines as: 0.98, 1.20, 1.44, 1.70 and 2.12 °C, respectively from 10 to 100 cm depth. This pattern is demonstrated in Figure 5 with respect to the RMSE. the prediction accuracy for the MLP-ANN also increases from 10 to 30 cm, then reduces from 50 to 100 cm.

Table 3. The performance evaluation results of both SVM-WAVE and MLP-ANN models

| Criterion | Model | Different Depths | | | | | |
|---|---|---|---|---|---|---|---|
| | | 5cm | 10cm | 20cm | 30cm | 50cm | 100cm |
| RMSE (°C) | SVM-WAVE | 2.49 | 1.22 | 1.54 | 1.78 | 2.07 | 2.47 |
| | MLP-ANN | 2.61 | 2.53 | 2.23 | 1.99 | 2.42 | 2.66 |
| MAE (°C) | SVM-WAVE | 1.27 | 0.98 | 1.2 | 1.44 | 1.7 | 2.12 |
| | MLP-ANN | 1.46 | 2.08 | 1.73 | 1.62 | 1.97 | 2.28 |
| MAPE (%) | SVM-WAVE | 16.83 | 9.26 | 7.88 | 8.23 | 8.5 | 9.12 |
| | MLP-ANN | 12.67 | 18.56 | 11.27 | 9.08 | 9.87 | 10.03 |
| APE (%) | SVM-WAVE | 5.89 | 4.45 | 5.47 | 6.34 | 7.12 | 8.64 |
| | MLP-ANN | 6.77 | 9.45 | 7.86 | 7.12 | 8.25 | 9.29 |
| NS (%) | SVM-WAVE | 96.03 | 98.99 | 97.98 | 96.65 | 94.33 | 85.21 |
| | MLP-ANN | 95.64 | 95.67 | 95.81 | 95.85 | 92.23 | 82.93 |

Both models produce the worst results at 100 cm. The coupling SVM-WAVE produces the best results at 10 cm, while the MLP-ANN best results are usually obtained at 30 cm depth. So, the designed hybrid SVM-WAVE is more compatible and more efficient with lower depths (i.e., 10 and 20 cm), while the MLP-ANN is more suitable for middle depths (i.e., 20 and 30 cm). In In depth 5 cm, unlike other depths, no peculiar pattern was observed. This may be due to sophistications of the soil temperature predictions at surface, since the ST changes significantly at surface compared to other depths (Mihalakakou, 2002). Figure 5 depicts the scatter plots of both models for all depths. Moreover, the trend line equations and coefficient determination $R^2$ corresponding to both SVM-WAVE (with darker font) and MLP-ANN (with brighter font) have been provided. As it is displayed in Figure 5 (a-f), the SVM-WAVE are closer to their relating trend line in comparison to MLP-ANN predictions. Also, the $R^2$ values for SVM-WAVE higher and closer to 1. The information provided in this figure confirms the previous clams the accuracy and error indices. For instance, the results of R2 indicate that the prediction accuracy of the SVM-WAVE decreases as the depth increases from 10 to 100 cm. The minimum and maximum differences in the obtained results by both models are 0.0024 and 0.0312 at depths 5 and 100 cm, respectively.

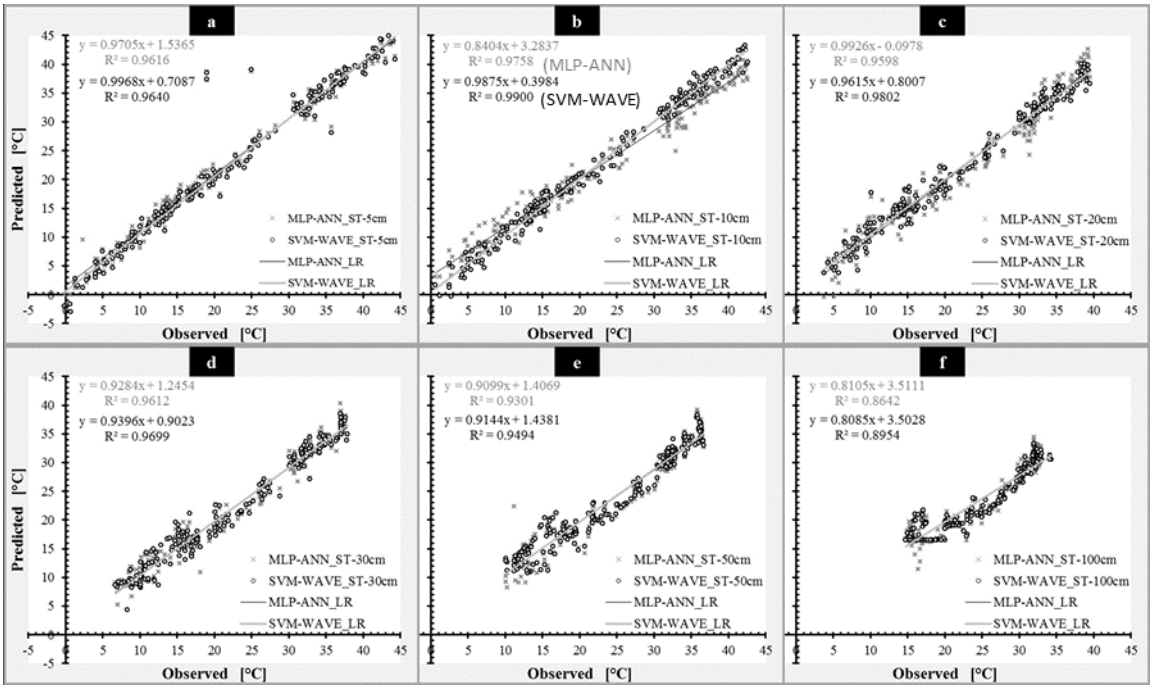


Figure 5. The scatter plots of ST predictions by both AI models, at depths (a) 5cm to (f) 100 cm

**CONCLUSIONS** Soil temperature as an crucial variable impacting human and wildlife, governs the crops life cycle, many microbial activities and consequently growth of plants. This parameter mutually affects climate change. In this study, an integrated Support Vector Machine (SVM) with Wavelet Transform (WT), SVM-WAVE, and the most conventional MLP-ANN were designed and structured for daily soil temperature prediction at depths of 5, 10, 20, 30, 50 and 100 cm. The daily measured meteorological data including 702 days (i.e., 23 months) have been obtained from Zahak synoptic meteorological station in southeast of Iran. The meteorological variables consisting of maximum air temperature ($T_{max}$), minimum air temperature ($T_{min}$), pan evaporation (EP), sunshine duration (SD) and soil temperature (ST). To compare models and evaluate their precision, RMSE, MAE, MAPE, NS, APE and $R^2$ indices were utilized. All the criteria indicated that the results obtained by the SVM-WAVE are more precise than those of the MLP-ANN. The SVM-WAVE by average indices of $R^2$ 0.953, RMSE 1.928, MAE 1.452, MAPE 9.970%,APE 6.318% and NS 94.865 outperformed the MLP-ANN by $R^2$ 0.947, RMSE 2.407, MAE 1.858, MAPE 11.913%, APE 8.123% and NS 93.022. Moreover, the results indicated that both models provide the worst results at depth 100 cm. Broadly the coupling SVM-WAVE produced its best results at 10 cm depth, but the MLP-ANN best results are usually obtained at depth 30 cm. Thus, the designed hybrid SVM-WAVE is more compatible and efficient with lower depths (i.e., 10 and 20 cm), while the MLP-ANN is more suitable and accurate with middle depths (i.e., 20, 30 and 50 cm).