




Daily soil temperature modeling using 'panel-data' concept

A. Mahabbati, A. Izady, M. Mousavi Baygi, K. Davary & S. M. Hasheminia


To cite this article: A. Mahabbati, A. Izady, M. Mousavi Baygi, K. Davary & S. M. Hasheminia (2016): Daily soil temperature modeling using 'panel-data' concept, Journal of Applied Statistics, DOI: [10.1080/02664763.2016.1214240](https://doi.org/10.1080/02664763.2016.1214240)

To link to this article: <http://dx.doi.org/10.1080/02664763.2016.1214240>

 View supplementary material 

 Published online: 16 Sep 2016.

 Submit your article to this journal 

 Article views: 23

 View related articles 

 View Crossmark data 

Daily soil temperature modeling using 'panel-data' concept

A. Mahabbati^a, A. Izady^b, M. Mousavi Baygi^a, K. Davary^a and S. M. Hasheminia^a

^aWater Engineering Dept., College of Agriculture, Ferdowsi University of Mashhad, Mashhad, Iran; ^bWater Research Center, Sultan Qaboos University, Muscat, Oman

ABSTRACT

The purpose of this research was to predict soil temperature profile using 'panel-data' models. Panel-data analysis endows regression analysis with both spatial and temporal dimensions. The spatial dimension pertains to a set of cross-sectional units of observation. The temporal dimension pertains to periodic observations of a set of variables characterizing these cross-sectional units over a particular time-span. This study was conducted in *Khorasan-Razavi* Province, Iran. Daily mean soil temperatures for 9 years (2001–2009), in 6 different depths (5, 10, 20, 30, 50 and 100 cm) under bare soil surface at 10 meteorological stations were used. The data were divided into two sub-sets for training (parameter training) over the period of 2001–2008, and validation over the period of the year 2009. The panel-data models were developed using the average air temperature and rainfall of the day before (T_{d-1} and R_{t-1} , respectively) and the average air temperature of the past 7 days (T_w) as inputs in order to predict the average soil temperature of the next day. The results showed that the two-way fixed effects models were superior. The performance indicators ($R^2 = 0.94$ to 0.99 , $RMSE = 0.46$ to 1.29 and $MBE = -0.83$ and 0.74) revealed the effectiveness of this model. In addition, these results were compared with the results of classic linear regression models using t -test, which showed the superiority of the panel-data models.

ARTICLE HISTORY

Received 21 May 2014
Accepted 14 July 2016


KEYWORDS

Soil temperature; panel-data model; linear regression; Iran; multivariate analysis

1. Introduction

Soil temperature is one of the most important meteorological factors in agricultural management [31], due to its great impact on plant growth. Temperature difference between the soil and the atmosphere is the primary driving force for soil water evaporation. Optimal temperatures are necessary for seed germination and normal growth of plants [23]. Furthermore, rate of most chemical reactions is affected by soil temperature profile [20]. However, soil temperature data in different depths is only available at meteorological stations, and all over the world, only few percent of the weather stations monitor it. Therefore, there would be a great interest for modeling or predicting soil temperature profile and its spatial variations.

CONTACT A. Mahabbati ✉ atbin.m@hotmail.com

 Supplemental data for this article can be accessed here. <http://dx.doi.org/10.1080/02664763.2016.1214240>

There are several methods for predicting the soil temperature, such as: analytical models [8,9,22], Fourier techniques [11,21], empirical equations (e.g. [30]), and artificial neural networks (ANNs) [10,34]. Although analytical models are accurate due to proven mathematical and physical background, they are inapplicable for practical purposes because of the size of the model and a lot of assumptions [35]. The problem with Fourier transform method is that its coefficients are just suitable for a particular site, which means they are not practical for simulations over many different sites [40]. ANN are greatly suited for dynamic nonlinear system modeling. However, these models tend to be used when understanding of the system is inadequate, and obtaining accurate predictions is more important than conceptualizing the actual physics of the system [7]. Although empirical models are simple and easy to use, they require large data bases from which to develop empirical coefficients for each specific site [27].

To look for a new method improving the modeling capabilities in this field, the main objective of this study was to investigate the possibility of 'Panel Data' concept [2,15] which seems to have the potential to predict soil temperature variability both spatially and temporally. Despite the vast application of the Panel-Data modeling in economies [13,16,24,28], its application in the field of environmental sciences is very young being initiated with the study of Izady *et al.* [17], who developed a Panel-Data-based model for predicting temporal fluctuations and spatial variations of groundwater level. The term 'panel-data' refers to the pooling of observations on a cross-section over several time periods. This can be achieved by surveying a number of observation sites or stations and following them over time. On the other hand, panel-data analysis endows regression analysis with both spatial and temporal dimensions. The spatial dimension pertains to a set of cross-sectional units of observations. The temporal dimension pertains to periodic observations of a set of variables characterizing these cross-sectional units over a particular time-span. The terms 'spatial' and 'cross-sectional' are used here in the sense of data, and not in the sense of physical landforms.

2. Materials and methods

2.1. Materials

This research was conducted in *Khorasan-Razavi* Province, north east of Iran, between 56°16'E to 61°16'E, and 33°23'N to 37°45'N (Figure 1). The climate of this area features a steppe climate (KoppenBSk) with hot summers and cool winters (KoppenBSK is the climate of a region that receives precipitation less than its potential evapotranspiration and is an intermediate between desert and humid climates in ecological and agricultural terminology). The maximum and minimum mean annual temperatures for the summer and winter seasons were 45.5 and 5.2, and 31.3 and -21.5°C, respectively. The average of annual air temperature was 15.7°C. The highest and lowest stations of the studied areas were located in *Torbat-Heidarieh* and *Sarakhs* with elevations of 1451 and 280 m above sea level, respectively. Daily mean soil temperatures for 9 years (2001–2009), in 6 different depths (5, 10, 20, 30, 50 and 100 cm) under the bare soil surface conditions at 10 meteorological stations were used. Mercury thermometers were used to measure the soil temperature. Soil thermometers were placed at 5, 10, 20, 30, 50 and 100 cm depths. An auger was used to dig holes for 50 and 100 cm depth soil thermometers, and they were placed such that to

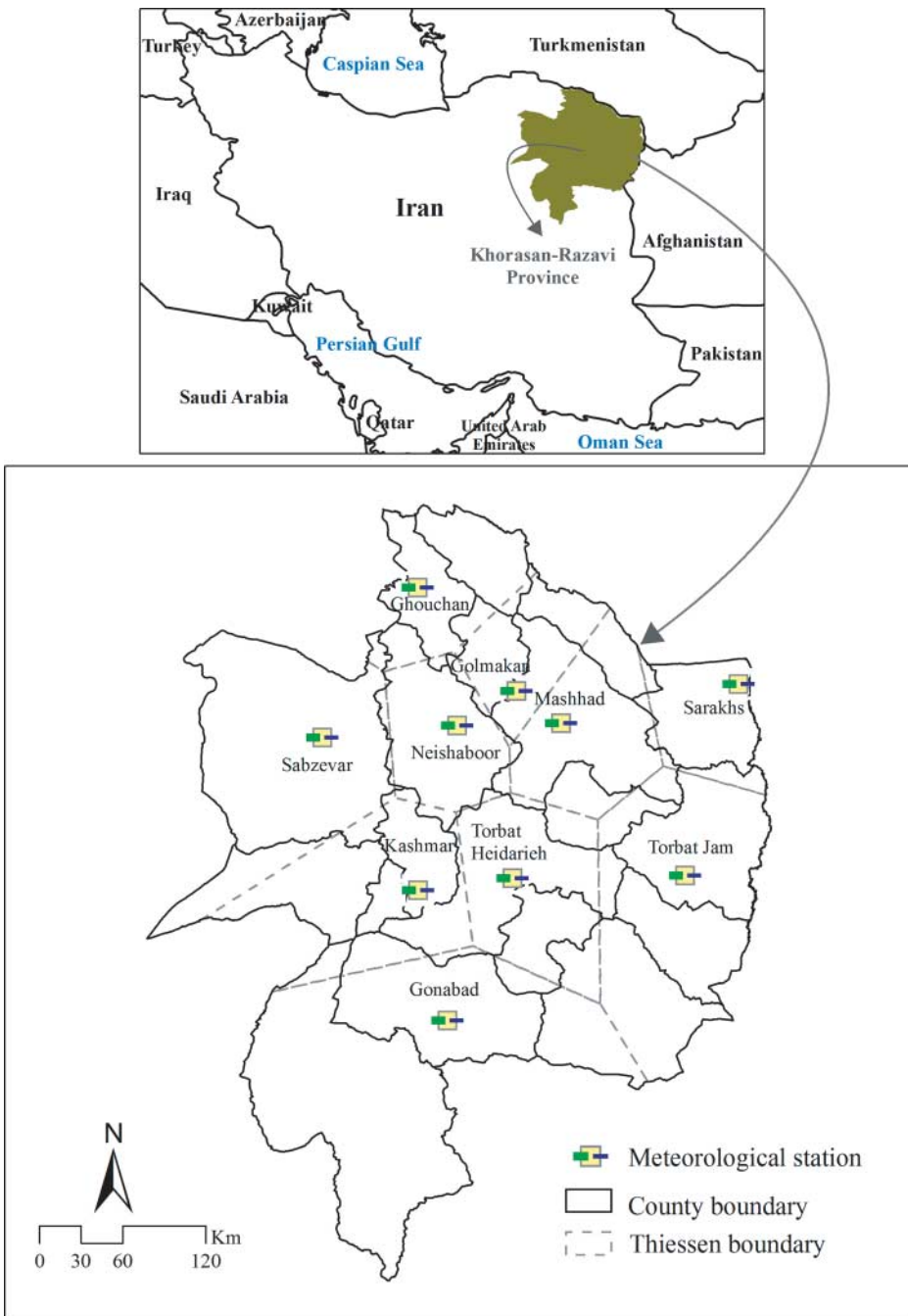


Figure 1. Location of study area in *Khorasan-Razavi* Province in north east of Iran.

have a good contact with the surrounding soil. The data were divided into two sub-sets for parameter training over the period of 2001 to 2008, and validation for the year of 2009. For more details about annual soil temperature data refer to Electronic Supplementary Material (ESM).

Soil temperature was measured three times daily (9:00, 12:00 and 15:00 GTM), and the average value was used for model development. Average daily air temperatures were obtained by calculating the average amount of minimum and maximum air temperatures for each day. The panel-data models were developed using air temperatures for one day and one week before in order to predict the average soil temperature of the next day. To capture the short and long-term effects of weather, average temperature of the day before (T_{d-1}) and the average air temperature of the past week (T_w) were used. For more details about air temperatures and rainfall amounts refer to ESM. The use of these parameters has been widely reported in the literature for soil temperature prediction [3,30,40]. This relation can be formulated in a panel-data model, as follows:

$$T_s = \alpha + \beta_1 T_{d-1} + \beta_2 T_w + \beta_3 R_{d-1} + \mu_i + \lambda_t, \quad (1)$$

where T_s is the average soil temperature at any Julian day ($^{\circ}\text{C}$), α is the general intercept ($^{\circ}\text{C}$), T_{d-1} is the daily air temperature ($^{\circ}\text{C}$) at Julian day $d-1$, T_w is the air temperature of the past week ($^{\circ}\text{C}$), R_{d-1} is the daily rainfall (mm) at Julian day $d-1$, μ_i and λ_t are unobservable *individual* and *time* effects, respectively, and β_1 , β_2 and β_3 are coefficients of independent variables.

In addition, classic linear regression (CLR) models were adopted to predict soil temperature using the same data. In these models, the relationship between the dependent and independent variables can be formulated as follows:

$$T_s = \alpha + b_1 T_{d-1} + b_2 T_w + b_3 R_{d-1}, \quad (2)$$

where b_1 , b_2 and b_3 are coefficients of independent variables. Inasmuch as temperature variations mechanisms are seasonally different due to seasonal effects [18], then it was decided to develop four seasonal separate models.

2.2. Theory of panel-data regression modeling

2.2.1. Introduction

The panel-data analysis is a kind of multivariate analysis which endows regression analysis with both a spatial and temporal dimension. The spatial dimension pertains to a set of cross-sectional units of observation. The temporal dimension pertains to periodic observations of a set of variables characterizing these cross-sectional units over a particular time-span. Such models can be viewed as follows [1,17,29,38,39]:

$$y_{it} = \alpha + \beta X_{it} + u_{it} \quad i = 1, 2, \dots, N; \quad t = 1, 2, \dots, T, \quad (3)$$

where i and t denotes the cross-section and time-series dimension, respectively, N is the number of cross-sections, T is the length of the time-series for each cross-section, y is a dependent-variable vector, X is an independent variable matrix, α is a scalar, β is the coefficient of the independent-variable matrix, and u is the error component in the model.

The performance of any estimation procedure for the model regression parameters depends on the statistical characteristics of the error components in the model. The panel-data procedure estimates the regression parameters in the preceding model under several common error structures. These error structures consist of *one* and *two-way fixed* and

random-effects models. If the specification is dependent only on the cross-section to which the observation belongs, such a model is referred to as a model with one-way effects. A specification that depends on both the cross section and the time-series to which the observation belongs is called as a model with two-way effects. Thus, the specifications for the one-way model are [2,15,38]:

$$u_{it} = \mu_i + v_{it}, \tag{4}$$

where μ_i denotes the *unobservable* individual-specific effect and v_{it} denotes the remainder disturbance. Note that μ_i is time-invariant and it accounts for any individual-specific effect that is not included in the regression. The remainder disturbance v_{it} varies with individuals and time and can be thought of as the usual disturbance in the regression. Similarly, the specifications for the two-way model are:

$$u_{it} = \mu_i + \lambda_t + v_{it}, \tag{5}$$

where λ_t denotes the unobservable time-specific effect. Note that λ_t is an individual-invariant and it accounts for any time-specific effect that is not included in the regression.

Apart from the possible one-way or two-way nature of the effects, the other dimension of difference between the possible specifications is due to the nature of the cross-sectional or time-series effect. The models are referred to as *fixed-effects* models if the effects are non-random, and as *random-effects* models otherwise [2,15,38].

2.2.2. The one-way fixed effects model

In this case, the μ_i are assumed to be fixed parameters to be estimated and the remainder disturbances stochastic with v_{it} independent and identically distributed $\text{IID}(0, \sigma_v^2)$. Note that σ_v^2 is the variance of the remainder disturbance. The X_{it} are assumed independent of the v_{it} for all i and t [2,12,15,19]. Afterwards, the ordinary least squares (OLS) estimator [25] is performed on Equation (3) to get estimates of α , β and μ . If N is large, then Equation (3) includes too many individual dummies, and the matrix to be inverted by OLS is large and of dimension $N + k$, where k is the number of independent variables. In fact, since α and β are the parameters of interest, the Least Squares Dummy Variables estimator can be obtained from Equation (3), by pre-multiplying the model by Q and performing OLS on the resulting transformed model ($Qy = QX\beta + Qv$) to get the coefficients. Note that Q is a matrix that obtains the deviations from individual means.

2.2.3. The one-way random effects model

In this case, $\mu_i \sim \text{IID}(0, \sigma_\mu^2)$, $v_{it} \sim \text{IID}(0, \sigma_v^2)$ and the μ_i are independent of the v_{it} . In addition, the X_{it} are independent of the μ_i and v_{it} , for all i and t . From Equation (3), the variance–covariance matrix of error can be computed as [2,15,38]:

$$\Omega = E(uu') = Z_\mu E(\mu\mu')Z_\mu' + E(vv'). \tag{6}$$

Note that Ω is variance–covariance matrix of error, $Z_\mu = I_N \otimes l_T$; where I_N is an identity matrix of dimension N , l_T is a vector of ones of dimension T , and \otimes denotes the Kronecker product [26,33]. Indeed, Z_μ is a selector matrix of ones and zeros, or simply the matrix of individual dummies that may be included in the regression to estimate the μ_i if those are assumed to be fixed parameters.

In order to obtain the generalized least square (GLS) estimator [5] of the regression coefficients, the Ω^{-1} is required. This is a huge matrix for typical panels and is of dimension $(NT + NT)$. After calculating Ω^{-1} using the method of Wansbeek and Kapteyn [36,37], GLS can be used as a weighted least-squares estimator to obtain coefficients for Equation (6).

2.2.4. The two-way fixed effects model

If the μ_i and λ_t are assumed to be fixed parameters to be estimated and the remainder disturbances stochastic with $v_{it} \sim \text{IID}(0, \sigma_v^2)$, then Equation (5) represents a two-way fixed effects error component model. The X_{it} are assumed independent of the v_{it} for all i and t . One would perform the regression of $\tilde{y} = Qy$ on $\tilde{X} = QX$ to get $\tilde{\beta}_{\text{OLS}} = (X'QX)^{-1}X'Qy$.

2.2.5. The two-way random effects model

If $\mu_i \sim \text{IID}(0, \sigma_\mu^2)$, $\lambda_t \sim \text{IID}(0, \sigma_\lambda^2)$ and $v_{it} \sim \text{IID}(0, \sigma_v^2)$ independent of each other, then this is the two-way random-effects model. In addition, X_{it} is independent of μ_i , λ_t and v_{it} for all i and t . From Equation (5), the variance-covariance matrix of error can be computed as follows [1,2,15,38]:

$$\Omega = E(uu') = Z_\mu E(\mu\mu')Z_\mu' + Z_\lambda E(\lambda\lambda')Z_\lambda' + \sigma_v^2 I_{NT}, \quad (7)$$

where Z_λ is the matrix of time dummies that may be included in the regression to estimate the λ_t , if they are fixed parameters and I_{NT} is an identity matrix of dimension NT . In order to obtain the GLS estimator of the regression coefficients, the Ω^{-1} is required. After calculating Ω^{-1} using a method developed by Hsiao [15], GLS can be used as a weighted least-squares estimator to obtain coefficients.

2.2.6. Fixed or random effects model

Having the fixed-effects and the random-effects models and their underlying assumptions discussed, the main question now arises that which one should be chosen. To answer this question, the following steps were taken. First, data poolability must be examined. The critical assumption behind pooling data into a panel is that the regression coefficients are constant across individuals (either all coefficients in the vector δ or at least the slope coefficients β). The pooled model, therefore, has constant coefficients. The Chow test [6] was used to examine data poolability, which is as follows:

H_0 : No individual fixed effects (the pooled model) ($\delta_1 = \delta_2 = \dots = \delta_N = \delta$)

H_1 : Individual fixed effects exist ($\delta_1 \neq \delta_2 \neq \dots \neq \delta_N$)

It is notable that the appropriate statistic for this hypothesis is the F -statistic:

$$F_{[(n-1)(k+1), n(T-(k+1))]} = \frac{(R_0^2 - R_1^2)/(n-1)(k+1)}{R_0^2/n(T-(k+1))}, \quad (8)$$

where R_0^2 is the sum square error (SSE) of the pooled model and R_1^2 is the SSE of the fixed effects model. If F is larger than a critical (tabulated) value, then the null hypothesis is rejected. It reveals the existence of fixed effects between unobservable individual-specific effects and regressors. After understanding the existent effect between individuals, it is necessary to find whether there are any random effects between individuals. With regard to this objective, different tests are proposed.

For the random two-way error-component model, Breusch and Pagan [4] suggested the Lagrange multiplier (LM) test. The assumptions are as follows:

H_0 : No random effects (the pooled model) ($\sigma_\mu^2 = \sigma_\lambda^2 = 0$)

H_1 : Random effects exist ($\sigma_\mu^2 > 0$ and $\sigma_\lambda^2 > 0$)

The LM test statistic is given by:

$$LM = LM_1 + LM_2 = \frac{nT}{2(T-1)} \left[1 - \frac{\tilde{u}'(I_N \otimes J_T)\tilde{u}}{\tilde{u}'\tilde{u}} \right]^2 + \frac{nT}{2(N-1)} \left[1 - 1 - \frac{\tilde{u}'(J_N \otimes I_T)\tilde{u}}{\tilde{u}'\tilde{u}} \right]^2, \tag{9}$$

where \tilde{u} is the SSE of the pooled model and J is a matrix of ones of dimension T or N . LM is asymptotically distributed as a χ^2 . If LM is larger than the critical value, then the null hypothesis is rejected. It means that there are random effects between unobservable individual-specific effects and regressors.

The Hausman specification test [14] is another classical test of whether the fixed or random effects model should be used. The main question here is whether there is significant correlation between the unobserved individual-specific random effects and the regressors. If there is no such correlation, then the random effects model may be more powerful. If there is such a correlation, the random effects model would be inconsistently estimated, and the fixed effects model would be the model of choice, which is as follows:

$H_0: E(X_{it}\mu_i) = 0 \rightarrow$ No correlation; random effects are consistent and efficient

$H_1: E(X_{it}\mu_i) \neq 0 \rightarrow$ Correlation exists; fixed effects are consistent

Hence, the Hausman test statistic is given by:

$$m = (\tilde{\beta}_{GLS} - \tilde{\beta}_{OLS})' [var(\tilde{\beta}_{GLS} - \tilde{\beta}_{OLS})]^{-1} (\tilde{\beta}_{GLS} - \tilde{\beta}_{OLS}). \tag{10}$$

The statistic m is asymptotically distributed as χ_k^2 where k denotes the number of regressors. If m is larger than the critical value, then the null hypothesis is rejected and the fixed effects model is selected. To implement the theory and to estimate or analyze panel-data models, StataSE software version 10 was used.

In summary, panel-data analysis is a method of studying a particular subject within multiple sites, periodically observed over a defined time frame. Moreover, with spatial observations and enough cross-sections, panel-data analysis permits the researcher to study the dynamics of change with time-series [17].

Different criteria were used in order to evaluate the effectiveness of the model and its ability to make proper predictions, as well as to compare the two models. These included coefficient of determination (R^2), root mean square error (RMSE), mean biased error (MBE), relative error (RE) and Akaike information criterion (AIC). The R^2 , RMSE and RE are well known and only the MBE and AIC coefficient are defined here:

$$MBE = \frac{\sum_{i=1}^n (x_i - y_i)}{n}, \tag{11}$$

where x and y are measured and estimated temperatures, respectively, and n is the number of observations.

$$AIC = 2k + n \log(RSS/n), \tag{12}$$

where K is the number of model parameters and n is the sample size and RSS is the Residual Sum of Squares.

3. Results and discussion

3.1. Model development

As mentioned earlier, the yearly data was separated into four seasons (I: from January to March, II: from April to June, III: from July to September and IV: from October to December). For these models, according to 6 depths and 4 seasons, 24 models were trained; where each model comprises all 10 stations. For each model, average air temperature and rainfall of the previous day and the average air temperature of the past week were considered as independent variables, and soil average daily temperatures in any depth were considered as the dependent variable. Firstly, the one-way and two-way fixed and random effect models were trained. Afterwards, Chow and Hausman tests were applied to determine the best model. First, the Chow test was performed for each model and showed that the fixed-effects models were superior. Then, the results of Hausman test illustrated that the two-way fixed-effects models were superior to the random-effects ones. Table 1 shows the results of the Chow and Hausman tests for the first panel-data model (n.b.: only results for 5 cm depth and first season are illustrated). The values were calculated using StataSE software version

Table 1. Computed values of the Chow and Hausman tests.

Test	Computed value (using Stata software)	Prob > F
Chow	153.37	0.001
Hausman	1401.87	0.001

Table 2. Number of observations, performance indices and parameters for winter in different depths.

Depth	Number of observations	R^2		RMSE		AIC	
		PD	CLR	PD	CLR	PD	CLR
5 cm	6300	0.97	0.84	1.01	2.38	10.6	12.9
10 cm	6300	0.97	0.87	0.85	1.96	10.2	12.5
20 cm	6300	0.97	0.88	0.81	1.72	9.9	12.4
30 cm	6300	0.97	0.87	0.77	1.60	9.8	12.3
50 cm	6300	0.97	0.81	0.64	1.57	8.0	9.9
100 cm	6300	0.96	0.50	0.53	1.78	7.7	10.4

Table 3. Number of observations, performance indices and parameters for the spring in different depths.

Depth	Number of observations	R^2		RMSE		AIC	
		PD	CLR	PD	CLR	PD	CLR
5 cm	6370	0.97	0.91	1.29	2.97	11.0	13.3
10 cm	6370	0.97	0.92	1.01	2.48	10.7	12.9
20 cm	6370	0.98	0.91	0.98	2.30	10.5	12.8
30 cm	6370	0.98	0.90	0.92	2.29	10.4	12.8
50 cm	6370	0.98	0.85	0.82	2.33	8.6	10.1
100 cm	6370	0.98	0.79	0.68	2.38	8.5	10.4

Table 4. Number of observations, performance indices and parameters for summer in different depths.

Depth	Number of observations	R^2		RMSE		AIC	
		PD	CLR	PD	CLR	PD	CLR
5 cm	6440	0.94	0.84	1.12	2.08	10.3	13.1
10 cm	6440	0.96	0.85	0.85	1.78	10.0	12.5
20 cm	6440	0.96	0.81	0.73	1.69	9.9	11.9
30 cm	6440	0.97	0.75	0.64	1.81	9.2	12.5
50 cm	6440	0.97	0.64	0.59	1.90	8.1	9.8
100 cm	6440	0.97	0.35	0.46	2.14	7.3	10

Table 5. Number of observations, performance indices and parameters for the autumn in different depths.

Depth	Number of observations	R^2		RMSE		AIC	
		PD	CLR	PD	CLR	PD	CLR
5 cm	6440	0.98	0.94	0.97	2.24	10.5	12.8
10 cm	6440	0.99	0.95	0.83	1.99	10.2	12.5
20 cm	6440	0.99	0.96	0.82	1.88	10.1	12.4
30 cm	6440	0.99	0.95	0.76	1.82	10.0	12.3
50 cm	6440	0.99	0.92	0.69	1.85	8.3	10.1
100 cm	6440	0.99	0.86	0.53	1.99	8.1	10.2

Table 6. The model error (RMSE and MBE) of panel-data for winter in different stations and depths in 2009.

Stations	5 cm		10 cm		20 cm		30 cm		50 cm		100 cm	
	RMSE	MBE	RMSE	MBE	RMSE	MBE	RMSE	MBE	RMSE	MBE	RMSE	MBE
Ghoochan	1.21	-0.13	1.09	-0.23	0.88	-0.44	0.86	-0.28	0.84	-0.52	0.75	-0.36
Golmakan	1.44	-0.06	1.33	-0.27	1.06	-0.33	0.82	-0.02	0.76	-0.07	0.66	0.31
Gonabad	1.53	0.39	1.21	0.02	1.30	0.37	1.13	0.12	0.96	0.41	0.83	0.29
Kashmar	1.25	-0.09	1.30	-0.01	1.16	-0.05	1.12	-0.35	0.91	0.12	0.96	-0.05
Mashhad	1.49	0.11	1.44	0.06	1.29	0.39	1.22	0.29	0.77	0.46	0.59	0.47
Neishabour	1.37	-0.07	1.32	-0.08	1.05	-0.14	0.89	0.18	0.71	0.27	0.65	0.23
Sabzevar	1.52	0.42	1.39	0.26	1.34	0.44	1.19	0.74	1.04	0.61	0.93	0.63
Sarakhs	1.49	0.12	1.36	0.14	1.32	0.73	1.13	0.43	0.92	0.34	0.87	0.65
Torbat H	1.42	-0.65	1.29	-0.67	1.29	-0.63	1.13	-0.67	0.86	-0.33	0.81	-0.07
Torbat J	1.41	0.00	1.38	-0.43	1.12	-0.49	0.99	-0.16	0.85	-0.26	0.68	-0.14

10. Therefore, the two-way fixed-effects model was opted for each depth and period as the most adequate model (24 models).

Tables 2–5 present performance indices and the number of observations of models for each period and depth, respectively. All parameters were found statistically significant at the level of $P < 0.01$ except R_{d-1} for depths of 50 cm and 100 cm for all models and T_{d-1} for the third season at the depths of 100 cm. The RMSE of models varied from 0.46 to 1.29 for PD and from 1.57 to 2.97 for CLR. The R^2 were significantly high and ranged between 0.94 and 0.99 for panel-data models, while these values were lower for CLR, ranging between 0.35 and 0.96. Note that R^2 and RMSE for CLR is calculated based on Thiessen area of meteorological stations. In fact, these indices were calculated for each station separately at a specified depth. The data presented in tables also show that in panel-data models, as the depth of soils increased, the RMSEs declined and coefficients of T_{d-1} and R_{d-1} decreased. It shows that the effect of daily air temperature was in inverse proportion to

Table 7. The model error (RMSE and MBE) of panel-data for spring in different stations and depths in 2009.

Stations	5 cm		10 cm		20 cm		30 cm		50 cm		100 cm	
	RMSE	MBE	RMSE	MBE	RMSE	MBE	RMSE	MBE	RMSE	MBE	RMSE	MBE
Ghoochan	1.26	-0.09	1.36	-0.23	1.19	-0.72	1.16	-0.83	1.02	-0.84	0.72	-0.45
Golmakan	1.39	-0.54	1.24	-0.56	1.23	-0.57	0.94	0.03	0.92	-0.48	0.81	-0.37
Gonabad	1.43	0.53	1.26	0.39	0.94	-0.22	0.88	-0.30	0.84	-0.51	0.86	-0.45
Kashmar	1.39	-0.68	1.18	0.15	1.05	0.17	0.91	-0.50	0.87	-0.18	0.86	0.51
Mashhad	1.26	-0.19	1.28	0.06	1.21	-0.69	0.98	0.51	0.94	0.47	0.92	0.38
Neishabour	1.44	0.10	1.34	0.35	1.29	-0.59	1.29	-0.63	1.24	-0.61	1.12	-0.52
Sabzevar	1.47	-0.79	1.36	-0.68	1.27	-0.57	1.19	-0.52	1.13	-0.58	0.89	-0.26
Sarakhs	1.47	0.41	1.19	0.10	1.17	0.68	1.14	0.46	1.01	0.51	0.94	0.36
Torbat H	1.61	-0.82	1.37	-0.77	1.31	-0.72	1.23	-0.59	1.13	-0.52	0.98	-0.42
Torbat J	1.52	-0.74	1.43	-0.73	1.37	-0.77	1.21	-0.65	1.18	-0.53	1.05	-0.49

Table 8. The model error (RMSE and MBE) of panel-data for summer in different stations and depths in 2009.

Stations	5 cm		10 cm		20 cm		30 cm		50 cm		100 cm	
	RMSE	MBE	RMSE	MBE	RMSE	MBE	RMSE	MBE	RMSE	MBE	RMSE	MBE
Ghoochan	1.13	-0.22	0.97	-0.58	0.93	-0.73	0.86	-0.32	0.84	-0.45	0.64	-0.21
Golmakan	1.21	-0.18	1.17	-0.38	1.01	-0.21	0.99	0.09	0.93	0.02	0.93	-0.38
Gonabad	1.33	0.43	1.13	0.20	0.85	-0.40	0.72	-0.07	0.67	0.27	0.57	-0.19
Kashmar	1.19	-0.66	1.15	-0.02	1.02	-0.22	1.12	-0.43	0.96	0.20	0.82	-0.01
Mashhad	1.28	-0.44	1.11	-0.27	1.07	0.26	1.02	0.39	0.94	0.46	0.94	0.49
Neishabour	1.29	-0.54	1.22	0.21	1.16	-0.56	1.14	-0.54	1.03	-0.38	0.98	-0.74
Sabzevar	1.44	-0.56	1.35	-0.49	1.21	-0.48	1.03	-0.55	0.95	-0.39	0.84	-0.30
Sarakhs	1.22	-0.34	1.18	-0.58	1.07	0.22	0.96	0.16	0.89	0.36	0.74	-0.02
Torbat H	1.46	-0.51	1.31	-0.55	1.29	-0.42	1.03	-0.62	0.95	-0.20	0.90	-0.68
Torbat J	1.25	-0.04	1.09	-0.51	1.09	-0.66	0.96	-0.32	0.88	-0.10	0.85	-0.47

Table 9. The model error (RMSE and MBE) of panel-data for autumn in different stations and depths in 2009.

Stations	5 cm		10 cm		20 cm		30 cm		50 cm		100 cm	
	RMSE	MBE	RMSE	MBE	RMSE	MBE	RMSE	MBE	RMSE	MBE	RMSE	MBE
Ghoochan	1.26	-0.53	1.24	-0.61	1.15	-0.64	0.92	-0.43	0.72	-0.49	0.69	-0.56
Golmakan	1.43	-0.46	1.22	-0.32	1.07	0.02	0.98	0.34	0.79	0.22	0.82	-0.17
Gonabad	1.33	-0.58	1.30	-0.64	1.24	-0.58	1.20	-0.52	1.16	-0.61	1.09	-0.48
Kashmar	1.13	-0.32	1.11	-0.56	1.09	-0.58	0.90	-0.52	0.73	0.06	0.79	-0.15
Mashhad	1.31	-0.72	1.26	-0.72	1.07	-0.39	0.82	0.10	0.77	-0.09	0.68	-0.12
Neishabour	1.25	-0.57	1.18	-0.61	1.13	-0.53	1.06	-0.22	1.01	-0.49	0.95	-0.52
Sabzevar	1.36	0.37	1.28	0.21	1.03	0.70	0.99	0.64	0.92	0.06	0.72	-0.55
Sarakhs	1.40	0.55	1.27	0.30	1.05	0.57	1.08	0.67	1.04	0.62	1.03	0.69
Torbat H	1.46	-0.69	1.33	-0.61	1.27	-0.72	1.19	-0.68	1.09	-0.59	1.02	-0.60
Torbat J	1.43	-0.60	1.34	-0.58	1.02	-0.54	1.02	0.00	0.95	-0.39	0.94	-0.59

depth; which seemed to be reasonable. This confirms the well-known relation between long-term air temperatures and deep soil temperatures. Moreover, AIC of models varied from 7.3 to 11 for PD and from 9.8 to 13.3 for CLR, which shows the superiority of PD compared with CLRs.

Table 10. The model RE of panel-data for winter in different stations and depths in 2009.

Stations	5 cm	10 cm	20 cm	30 cm	50 cm	100 cm
	RE	RE	RE	RE	RE	RE
Ghoochan	0.082	0.077	0.077	0.083	0.098	0.125
Golmakan	0.089	0.087	0.082	0.078	0.084	0.094
Gonabad	0.071	0.062	0.081	0.078	0.085	0.122
Kashmar	0.057	0.062	0.065	0.071	0.072	0.133
Mashhad	0.077	0.079	0.086	0.087	0.081	0.120
Neishabour	0.070	0.071	0.080	0.080	0.080	0.105
Sabzevar	0.080	0.087	0.110	0.116	0.58	0.213
Sarakhs	0.063	0.065	0.076	0.073	0.072	0.141
Torbat H	0.079	0.077	0.084	0.082	0.088	0.148
Torbat J	0.069	0.074	0.077	0.082	0.087	0.116

Table 11. The model RE of panel-data for spring in different stations and depths in 2009.

Stations	5 cm	10 cm	20 cm	30 cm	50 cm	100 cm
	RE	RE	RE	RE	RE	RE
Ghoochan	0.057	0.064	0.063	0.064	0.065	0.060
Golmakan	0.053	0.048	0.054	0.043	0.052	0.062
Gonabad	0.048	0.047	0.042	0.039	0.046	0.064
Kashmar	0.065	0.056	0.052	0.045	0.048	0.055
Mashhad	0.056	0.063	0.067	0.054	0.059	0.067
Neishabour	0.048	0.050	0.063	0.065	0.073	0.089
Sabzevar	0.055	0.059	0.062	0.060	0.066	0.070
Sarakhs	0.056	0.054	0.056	0.056	0.055	0.061
Torbat H	0.095	0.087	0.089	0.088	0.089	0.094
Torbat J	0.056	0.060	0.068	0.062	0.072	0.083

Table 12. The model RE of panel-data for summer in different stations and depths in 2009.

Stations	5 cm	10 cm	20 cm	30 cm	50 cm	100 cm
	RE	RE	RE	RE	RE	RE
Ghoochan	0.096	0.088	0.092	0.101	0.131	0.162
Golmakan	0.071	0.095	0.085	0.094	0.103	0.185
Gonabad	0.102	0.096	0.093	0.093	0.093	0.259
Kashmar	0.075	0.073	0.077	0.096	0.098	0.138
Mashhad	0.067	0.071	0.087	0.093	0.116	0.203
Neishabour	0.077	0.116	0.174	0.163	0.172	0.294
Sabzevar	0.124	0.139	0.173	0.149	0.180	0.257
Sarakhs	0.076	0.090	0.092	0.094	0.103	0.142
Torbat H	0.116	0.124	0.151	0.140	0.166	0.250
Torbat J	0.094	0.100	0.140	0.138	0.162	0.252

3.2. Test of the models

The performances of the panel-data and CLR models were tested by comparing models calculations with observed data of 2009 at the 10 meteorological stations. The differences between predicted temperatures and measured ones calculated for all stations and subsequently their RMSE and MBE were calculated and presented for panel-data models in Tables 6–9. Furthermore, related errors of panel-data models were also calculated and presented in Tables 10–13.

Figure 2 shows the differences between the averages of measured and estimated amounts of soil temperatures during the year 2009 for both panel-data and CLR models. According

Table 13. The model RE of panel-data for autumn in different stations and depths in 2009.

Stations	5 cm	10 cm	20 cm	30 cm	50 cm	100 cm
	RE	RE	RE	RE	RE	RE
Ghoochan	0.058	0.059	0.059	0.051	0.041	0.046
Golmakan	0.057	0.052	0.051	0.051	0.047	0.062
Gonabad	0.047	0.048	0.049	0.050	0.054	0.066
Kashmar	0.042	0.042	0.046	0.039	0.036	0.050
Mashhad	0.050	0.051	0.047	0.040	0.042	0.052
Neishabour	0.051	0.052	0.056	0.055	0.058	0.069
Sabzevar	0.059	0.060	0.056	0.057	0.057	0.053
Sarakhs	0.055	0.052	0.045	0.051	0.054	0.072
Torbat H	0.060	0.058	0.061	0.061	0.063	0.074
Torbat J	0.056	0.055	0.050	0.058	0.057	0.072

to this figure, the mean differences in most periods, were scattered in the range of -1 to 1°C for panel-data models (which seems to be acceptable according to similar researches (e.g. [30,32]) and from -1.5 to 1.5°C for CLR ones. The patterns of differences from measured temperatures were similar for panel-data and CLR models except in the depths of 50 and 100 cm, where panel-data models had a considerably greater performance. Moreover, during the Julian days of 110–140, all models overestimated the soil temperatures continuously in all depths. On the other hand, as Tables 6–9 indicate, RMSEs declined by depths in almost all stations and all periods in panel-data models. This could be due to the fact that the variability of temperature declines with depth. Nevertheless, based on Tables 10–13, REs did not show a significant change based on depth except at the depth of 100 cm in which REs were almost slightly higher than other depths. Furthermore, RMSEs varied from 0.57 to 1.61 and MBEs varied in the range of -0.83 to 0.74. Despite the significant RMSE patterns, MBEs were depth-independent. The lower RMSEs happened in *Ghoochan* station which can be explained by the fact that this station had the closest range of fluctuations in its annual air temperature regime. The closest average of MBEs to zero was occurred in *Mashhad* station because the characteristics of this station (height, average temperature and rainfall) was the closest one to the average of all 10 stations. In contrast, the biggest RMSEs and MBEs were occurred in the station of *Torbat-Heydarieh*. This can be justified by the fact that this station had the widest range of fluctuations in its annual air temperature regime and also it is located in the boundary of *Khorasan-Razavi* Province which is near the dryer and hotter southern area. In addition, the mean of residuals were different from zero in most of the stations and for most of the depths. In General the means of residuals which their MBE were less than -0.3 or more than 0.3 were significantly different from zero and it was depth independent. It should also be mentioned that none of the stations had biased residuals (consistently positive or consistently negative), while all of them were distributed normally. It should be noted that residuals do not demonstrate correlation.

3.3. Comparison of CLR and panel-data models

The results of panel-data models were compared with those of CLR ones using RMSE and MBE. According to Tables 14–17, RMSEs showed that panel-data models had better performances in all periods and all depths which can be explained by the ability of panel-data

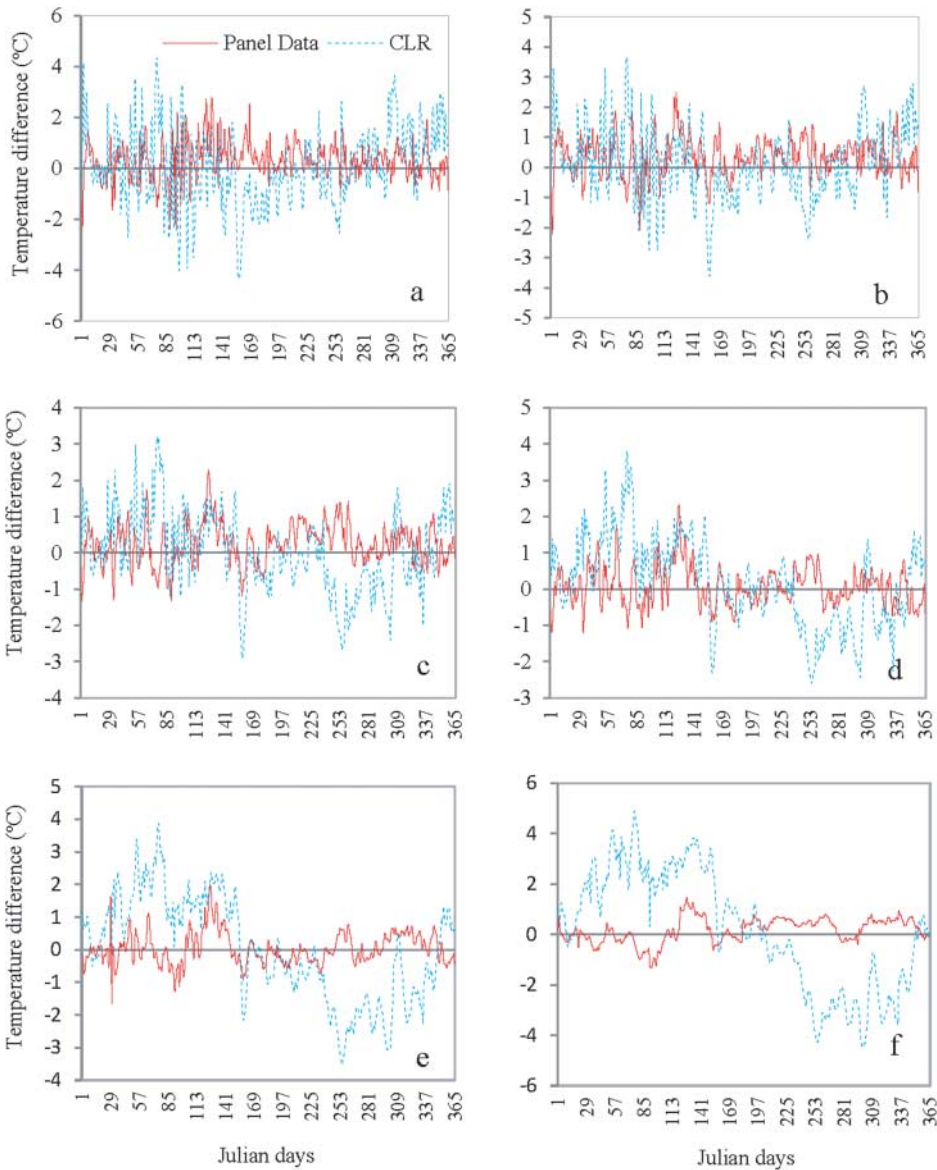


Figure 2. Differences between the average of measured and estimated amounts of soil temperatures during the year 2009 (°C) at the depth of (a) 5 cm, (b) 10 cm, (c) 20 cm, (d) 30 cm, (e) 50 cm and (f) 100 cm considering all stations.

models to include time effects. Furthermore, as shown in Tables 14–17, although the average of MBEs did not show any supremacy of neither panel-data nor CLR performances, CLR models were the ones which experienced the extreme amounts of MBEs. The RMSEs declined constantly from shallower depths – 5 cm – to deeper ones – 100 cm – in PD models; whereas, deeper depths – 50 and 100 cm – experienced some rises in CLR models. The minimum average of RMSEs occurred during summer for both models, while

Table 14. Validation indices for winter in different depths based on Thiessen weighted average of stations.

Depth	RMSE		MBE	
	PD	CLR	PD	CLR
5 cm	1.43	2.19	0.04	-0.34
10 cm	1.30	1.79	-0.12	-0.57
20 cm	1.22	1.75	0.02	-0.84
30 cm	1.09	1.94	0.02	-1.25
50 cm	0.89	2.13	0.15	-1.47
100 cm	0.80	2.78	0.20	-2.37
Average	1.12	2.10	0.05	-1.14

Table 15. Validation indices for spring in different depths based on Thiessen weighted average of stations.

Depth	RMSE		MBE	
	PD	CLR	PD	CLR
5 cm	1.44	2.66	-0.27	0.81
10 cm	1.31	2.13	-0.19	0.48
20 cm	1.17	1.60	-0.42	-0.10
30 cm	1.07	1.67	-0.37	-0.57
50 cm	1.01	1.82	-0.40	-0.88
100 cm	0.91	2.51	-0.23	-2.19
Average	1.15	2.07	-0.31	-0.41

Table 16. Validation indices for summer in different depths based on Thiessen weighted average of stations.

Depth	RMSE		MBE	
	PD	CLR	PD	CLR
5 cm	1.31	1.98	-0.23	0.31
10 cm	1.18	1.80	-0.25	0.22
20 cm	1.06	1.58	-0.36	0.30
30 cm	0.95	1.58	-0.27	0.31
50 cm	0.87	1.91	-0.01	0.87
100 cm	0.79	2.05	-0.27	0.95
Average	1.03	1.82	-0.23	0.49

the maximum ones happened during spring and winter seasons for PD and CLR models, respectively.

According to Table 18, the t -tests illustrated that the null hypotheses of panel-data RMSE average was equal to or more than CLR RMSE average rejected by t -tests; hence, the panel-data RMSEs were significantly less than the CLR ones. Nevertheless, t -tests did not prove any supremacy of neither panel-data nor CLR performances when the averages of MBE were compared.

Table 17. Validation indices for autumn in different depths based on Thiessen weighted average of stations.

Depth	RMSE		MBE	
	PD	CLR	PD	CLR
5 cm	1.34	2.06	-0.38	-1.10
10 cm	1.27	1.76	-0.44	-0.68
20 cm	1.14	1.73	-0.32	0.04
30 cm	1.05	1.70	-0.16	0.42
50 cm	0.96	1.90	-0.27	1.11
100 cm	0.90	2.58	-0.40	1.72
Average	1.11	1.96	-0.33	0.25

Table 18. A comparison between averages of RMSEs and MBEs of PD and CLR models using the data of Tables 10–13.

Factors	Averages of RMSE	Averages of MBE
<i>t</i> -Ratio	10.37	0.0201
DF	23	23
Prob > <i>t</i>	< 0.0001	0.9841
Prob > <i>t</i>	< 0.0001	0.4921
Prob < <i>t</i>	1.0000	0.5079

4. Conclusions

Panel-data models showed RMSEs from 0.46 to 1.29°C, considerably lower than those of CLR. Also, the averages of the R^2 for each season and depth were acceptable for PD models – varied between 0.94 and 0.99 – which were significantly greater than those of CLR, which were between 0.35 and 0.96. The PD models could predict soil temperatures at various depths and stations with mean errors in the range of -1 to 1°C for most of the year. Nonetheless, in some days during the second season (April and May) mean errors were larger, there was a constant overestimation of up to 2°C , especially in shallower depths. The overestimation of soil temperatures in the second season can be explained by the fact that the rainfall was significantly higher (more than twice) than the normal almost all over *Khorasan-Razavi* during spring, summer and autumn (especially during April and May) in 2009. In spite of the mean errors were found at a reasonable level, the absolute errors on certain locations and in shallower depths might, occasionally, have been in the order of 2.7°C , which can be explained by the higher fluctuations in the depths of 5 and 10 cm. In addition, the validation indicated that the panel-data models were useful and reliable for prediction of soil temperature; nevertheless, the effect of heavy and unusual rainfalls could lead to overestimation. Moreover, it should be noted that the accuracy of predictions improved by increasing soil depth. In contrast, it should be mentioned that in CLR models, the RMSEs did not follow the same pattern by increase in soil depth (while the same independent variables were used). Moreover, the RMSEs were substantially lower for panel-data models compared with CLR ones. Consequently, it seems that panel-data models can predict variables with more sinusoidal and organized patterns than those which have delay effects, much better than CLR ones.

Future investigations on application of panel-data to soil temperature modeling may comprehensively find pros and cons of panel-data approach in comparison with other

methods. Further studies are encouraged to examine the potentials of the panel-data concept for a broader usage and modeling capability in meteorology and environmental sciences.

Acknowledgements

The authors would like to thank Dr Majid Sarmad and Ms Zahra Khoshkam from Department of Statistics of Ferdowsi University of Mashhad for their insightful suggestions that led to a substantial improvement of the manuscript.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

- [1] M. Arellano, *Panel Data Econometrics*, 2nd ed., Oxford University Press, New York, 2003.
- [2] B.H. Baltagi, *Econometric Analysis of Panel Data*, 3rd ed., John Wiley & Sons, New York, 2005.
- [3] B. Bond-Lamberty, C., Wang, and S.T. Gover, *Spatiotemporal measurement and modeling of stand-level boreal forest soil temperatures*, *Agr. Forest Meteorol.* 131 (2005), pp. 27–40.
- [4] T.S. Breusch and A.R. Pagan, *The Lagrange multiplier test and its applications to model specification in econometrics*, *Rev. Econom. Stud.* 47 (1980), pp. 239–253.
- [5] M.W. Browne, *Generalized least squares estimators in the analysis of covariance structures*, *South African Statist. J.* 8 (1974), pp. 1–24.
- [6] G.C. Chow, *Tests of equality between sets of coefficients in two linear regressions*, *Econometrica* 28 (1960), pp. 591–605.
- [7] I.N. Daliakopoulos, P. Coulibaly, and I.K. Tsanis, *Groundwater level forecasting using artificial neural networks*, *J. Hydrol.* 309 (2005), pp. 229–240.
- [8] F. Droulia, S. Lykoudis, I. Tsiros, N. Alvertos, E. Akylas, and I. Garofalakis, *Ground temperature estimations using simplified analytical and semi-empirical approaches*, *Solar Energy* 83 (2009), pp. 211–219.
- [9] Z. Gao, L. Bian, Y. Hu, L. Wang, and J. Fan, *Determination of soil temperature in an arid region*, *J. Arid Environ.* 71 (2007), pp. 157–168.
- [10] R.K. George, *Prediction of soil temperature by using artificial neural network algorithms*, *Nonlinear Anal.* 47 (2001), pp. 1737–1748.
- [11] E.A. Graham, Y. Lam, and E.M. Yuen, *Forest understory soil temperatures and heat flux calculated using a Fourier model and scaled using a digital camera*, *Agric. Forest Meteorol.* 150 (2010), pp. 640–649.
- [12] B.H. Hall, *The relationship between firm size and firm growth in the US manufacturing sector*, *J. Ind. Econ.* 35 (1987), pp. 583–606.
- [13] M. Harding and C. Lamarche, *Least square estimation of a panel data model with multifactor error structure and endogenous covariates*, *Econom. Lett.* 111 (2011), pp. 197–199.
- [14] J. A. Hausman, *Specification tests in econometrics*, *Econometrica* 46 (1978), pp. 1251–1271.
- [15] C. Hsiao, *Analysis of Panel Data*, 2nd ed., Cambridge University Press, London, 2003.
- [16] C. Hsiao, M. Pesaran, and A. Tahmiscioglu, *Maximum likelihood estimation of fixed effects dynamic panel data models covering short time periods*, *J. Econometrics* 109 (2002), pp. 107–150.
- [17] A. Izady, K. Davary, A. Alizadeh, B. Ghahraman, M. Sadeghi, and A. Moghaddamnia, *Application of 'panel-data' modeling to predict groundwater levels in the Neishaboor Plain, Iran*, *Hydrogeol. J.* 20 (2012), 435–447. doi:10.1007/s10040-011-0814-2.
- [18] A. Jebamalar, S. Raja, and S. Bai, *Prediction of annual and seasonal soil temperature variation using artificial neural network*, *Indian J. Radio Space Phys.* 41 (2012), pp. 48–57.
- [19] A. Kangasharju, *Regional variations in firm formation: Panel and cross-section data evidence from Finland*, *Pap. Reg. Sci.* 79 (2000), pp. 355–373.

- [20] D. Kirkham and W.L. Powers, *Advanced Soil Physics*, first ed. John Wiley & Sons, New York, 1972.
- [21] A. Krishnan and R.S. Kushwaha, *Analysis of soil temperatures in the arid zone of India by Fourier techniques*, *Agric. Meteorol.* 10 (1972), pp. 55–64.
- [22] P. Kumar and A. Kaleita, *Assimilation of near-surface temperature using extended Kalman filter*, *Adv. Water Sources* 26 (2003), pp. 79–93.
- [23] R. Lal and M.K. Shukla, *Principles of Soil Physics*, 1st ed., Marcel Dekker, New York, 2004.
- [24] L. Lee and J. Yu, *Some recent developments in spatial panel data models*, *Reg. Sci. Urban Econ.* 40 (2010), pp. 255–271.
- [25] L. Leng, T. Zhang, L. Kleinman, and W. Zhu, *Ordinary least square regression, orthogonal regression, geometric mean regression and their applications in aerosol science*, *J. Phys.* (2007), Conference Series 78 012084.
- [26] S. Liu, *Matrix results on the Khatri-Rao and Tracy-Singh products*, *Linear Algebr. Appl.* 289 (1999), pp. 267–277.
- [27] Y. Luo, R.S. Loomis, and T.C. Hsiao, *Simulation of soil temperature in crops*, *Agric. Forest Meteorol.* 61 (1992), pp. 23–38.
- [28] M. Mouchart and J. Rombouts, *Clustered panel data models: An efficient approach for nowcasting from poor data*, *Int. J. Forecast.* 21 (2005), pp. 577–594.
- [29] Y. Mundlak, *On the pooling of time series and cross section data*, *Econometrica* 46 (1978), pp. 69–85.
- [30] F. Plauborg, *Simple model for 10 cm soil temperature in different soils with short grass*, *Eur. J. Agron.* 17 (2002), pp. 173–179.
- [31] N. Rosenberg, *Microclimate: The Biological Environment*, 1st ed., John Wiley & Sons, New York, 1974.
- [32] D.J. Timlin, Y.A. Pachepsky, B.A. Acock, J. Simunek, G. Flerchinger, and F. Whisler, *Error analysis of soil temperature using measured and estimated hourly weather data with 2DSOIL*, *Agric. Syst.* 72 (2002), pp. 215–239.
- [33] G. Trenkler, *A Kronecker matrix inequality with a statistical application*, *Econom. Theory* 11 (1995), pp. 654–655.
- [34] M.R. Veronez, A.B. Thum, A.S. Luz, and D.R. Dasilva, *Artificial neural networks applied in the determination of soil surface temperature-SST*, 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, 2006.
- [35] J. Wang and R.L. Bras, *Ground heat flux estimated from surface soil temperature*, *J. Hydrol.* 216 (1999), pp. 214–226.
- [36] T.J. Wansbeek and A. Kapteyn, *A simple way to obtain the spectral decomposition of variance components models for balanced data*, *Commun. Statist.* A11 (1982), pp. 2105–2112.
- [37] T.J. Wansbeek and A. Kapteyn, *A note on spectral decomposition and maximum likelihood estimation of ANOVA models with balanced data*, *Statist. Probab. Lett.* 1 (1983), pp. 213–215.
- [38] J.M. Wooldridge, *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge, 2002.
- [39] R. Yaffee, *A primer for panel data analysis*. Available via DIALOG, 2003. Available at http://www.nyu.edu/its/pubs/connect/fall03/yaffee_primer.html (accessed 15 April 2008).
- [40] D. Zheng, E.R. Hunt, and S.W. Running, *A daily soil temperature model based on air temperature and precipitation for continental applications*, *Climate Res.* 2 (1993), pp. 183–191.