MANUSCRIPT

A guideline to select an estimation model of daily global solar radiation between geostatistical interpolation and stochastic simulation approaches

D. I. Jeong[1), A. St-Hilaire[2), Y. Gratton[2), C. Bélanger[2), C. Saad[3)

[1]Centre ESCER (Étude et Simulation du Climat à l'Échelle Régionale), Université du Québec à Montréal, Montreal, Canada

[2]INRS-ETE, University of Québec, Québec, Canada

[3]Environment and Climate Change Canada, Montreal, Canada

Corresponding author address: Dae Il Jeong, Ph.D., P.Eng., Centre ESCER, Université du Québec à Montréal, 201 Ave. President-Kennedy, Montreal, Quebec H3C 3P8, Canada

Email: jeong@sca.uqam.ca

Abstract

This study compares geostatistical interpolation and stochastic simulation approaches for the estimation of daily global solar radiation (GSR) on a horizontal surface in order to fill in missing values and to extend short record length of a meteorological station. A guideline to select an approach is suggested based on this comparison. Three geostatistical interpolation models are developed using the nearest neighbor (NN), inverse distance weighted (IDW), and ordinary kriging (OK) schemes. Three stochastic simulation models are also developed using the artificial neural network (ANN) method with daily temperature (ANN(T)), relative humidity (ANN(H)), and both (ANN(TH)) variables as predictors. The six models are compared at 13 meteorological stations located across southern Quebec, Canada. The three geostatistical interpolation models yield better performances at stations located in a high density area of GSR measuring stations compared to the three stochastic simulation models. The guideline suggests an optimal approach by comparing a threshold distance, estimated according to a performance criteria of a stochastic simulation model, to the distance between a target and its nearest neighboring station. Additionally, the spatial correlation strength of daily GSRs and the at-site correlation strength between daily GSRs and the predictor variables should be considered.

Keywords: artificial neural networks, geostatistical interpolation, global solar radiation, spatial correlation, temperature, relative humidity.

1. Introduction

Global solar radiation (GSR) on a horizontal surface of the earth is an important variable for many analyses involving agricultural and plant growth, air and water temperatures, environmental and biological risk, and solar electric generation. However, instruments measuring solar irradiation (i.e., Kipp or Eppley pyranometers) are relatively expensive and difficult to manage [1], compared to those of common meteorological variables such as air temperature, precipitation, and relative humidity. Therefore, meteorological stations for GSR are generally less abundant than those for the common meteorological variables. Furthermore, observed GSR datasets are usually short timeseries and have large gaps of missing values.

Geostatistical interpolation approaches can be adopted to fill in missing values and to extend short record length of the GSR at a station using observed GSR data on the other stations located near the desired station. Kriging [2-7], nearest neighbor [4], and inverse distance weighted average [5,8,9] approaches have been applied frequently for the spatial interpolation.

At-site physical and statistical approaches can also be used for GSR simulations. Physical models (e.g., [10-12]) use complex physical interactions between the GSR and the terrestrial atmosphere, such as the Rayleigh scattering, radiative absorption by ozone and water vapour and aerosol extinction. Stochastic simulation models (e.g., [13-20]) use empirical relationships between GSR and meteorological covariables such as sunshine hours, temperature, and relative humidity at a desired station. This study considers stochastic simulation models as they are relatively simple to develop and require fewer input variables compared to physical models [16,17]. Although linear and non-linear regressions as well as artificial neural networks (ANNs) can be employed to drive empirical relationships between the common meteorological variables and the GSR, many studies [16-18,20-22] have shown the superiority of ANN approaches to regression-based approaches.

Sunshine duration, one of the most explanatory variables for GSR simulation [18,21,23], has not been recorded at most meteorological stations in Canada since 1999 due to its difficulty of measurement [1]. Temperature [13-15,17-19,24-28] and relative humidity [18,21] are alternative covariables although they have weaker correlation with GSR compared to sunshine duration [18].

Geostatistical interpolation and statistical simulation approaches for GSR estimation have been applied separately in many studies, however, they have been rarely compared in an application study. Therefore, this study compares geostatistical interpolation and statistical simulation approaches to fill in missing values and to extend short record length of daily GSR timeseries. The spatial interpolation approaches considered include the nearest neighbor, the inverse distance weighted, and the ordinary kriging methods. The stochastic simulation models include three ANN-based models with daily temperature and/or daily relative humidity as input variables. The six models are applied at 13 meteorological stations located across southern Quebec (45.1~50.3 °N and 64.2~79.0 °W), Canada. Furthermore, a guideline to choose an approach between the geostatistical interpolation and the statistical simulation approaches is provided for the estimation of daily GSR on the study area.

2. Methodologies

2.1 Geostatistical interpolation models

Three geostatistical interpolation models are developed based on nearest neighbor (NN), inverse distance weighted (IDW), and ordinary kriging (OK) schemes for daily GSR. The NN model employs the simplest algorithm among the three models. This model selects the value of the nearest station to the location of interest and does not consider the other values of neighboring stations in order to yield a piecewise-constant interpolation map.

The IDW interpolation algorithm adopts the assumption that the interpolation value at a location of interest is inversely proportional to the distances of nearby stations. The interpolation value of the model is a weighted average of the values of multiple stations and the weight assigned to each nearby station diminishes as the distance from the interpolation point to that station increases. The IDW model interpolates the daily GSR value $R(x_0)$ at an ungauged location x_0 from observations $R(x_i)$ at locations $x_1, ..., x_n$ as follows:

$$\hat{R}(x_0) = \sum_{i=1}^{n} w_i R(x_i) \tag{1}$$

$$w_i = \frac{1/d_i}{\sum_{j=1}^{n} 1/d_j}, \quad i = 1, 2, ..., n \tag{2}$$

where $\hat{R}(x_0)$ is an interpolated value of $R(x_0)$ and d_i represents distance between $R(x_0)$ and $R(x_i)$.

Kriging is a geostatistical interpolation technique based on the linear least square estimation algorithm. Ordinary kriging (OK) is the most common among many kriging approaches. OK estimates the best linear unbiased estimator based on a linear model. The interpolation value of the OK at a location x_0 is given by the following equation:

$$\hat{R}(x_0) = \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} R(x_1) \\ \vdots \\ R(x_n) \end{pmatrix} \tag{3}$$

where $w_1, ..., w_n$ are the weights of the OK that fulfill the unbiased condition $\sum_{i=1}^{n} w_i = 1$. The weights are obtained by the below OK equation system:

$$\begin{pmatrix} w_1 \\ \vdots \\ w_n \\ \mu \end{pmatrix} = \begin{pmatrix} \gamma(x_1, x_1) & \cdots & \gamma(x_1, x_n) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \gamma(x_n, x_1) & \ldots & \gamma(x_n, x_n) & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix}^{-1} \begin{pmatrix} \gamma(x_1, x_0) \\ \vdots \\ \gamma(x_n, x_0) \\ 1 \end{pmatrix} \tag{4}$$

where $\mu = E[R(x)]$ is a Lagrange parameter employed to minimize the kriging error under the unbias condition, which is assumed to be an unknown constant in the OK. $\gamma(x_i, x_j)$ is a variogram function to calculate the spatial dependency between $R(x_i)$ and $R(x_j)$. Several variogram functions are available such as exponential, Gaussian, and spherical models. In this study, the spherical variogram function is selected based on trial and error examination. The variogram is estimated for each day, based on observed daily GSR dataset of nearby stations. The detail descriptions of variogram models and ordinary kriging can be found in [29,30].

To verify the interpolation performances of the three models, a leave-one-out cross-validation approach is employed. Among the observations at *n* stations, GSR values of one of those stations are interpolated using the observations at the remaining *n*-1 stations. This process is repeated for all the observation stations. The interpolated $\hat{R}(x_i)$ is then compared to the associated observation $R(x_i)$ at each station in order to evaluate the performance of the interpolation models.

2.2 Stochastic simulation models

Three stochastic simulation models are developed to estimate daily GSR using the ANN approach as a transfer function and daily maximum and minimum temperatures and/or daily mean relative humidity as input variables. Feed forward ANNs have been frequently employed to simulate GSR [16-18, 20, 21] from the meteorological input variables. This study also employs a three-layer feed forward ANN model, which includes an input layer, a single hidden layer, and an output layer of computation nodes. The ANN models are trained by the Bayesian regularization backpropagation (BRBP) algorithm, which is a network training function that updates the weight and bias values according to the Levenberg-Marquardt optimization [31]. An important issue in ANN modelling is the determination of the number of hidden nodes. Fletcher and Goss [32] suggested that the optimal number of hidden nodes could be within ($2p^{0.5}+ o$) ~ ($2p$+1), where p and o are the numbers of independent and dependent variables, respectively. The hyperbolic tangent sigmoid function is employed for the hidden layer and the linear function is used for the output layer. Detailed descriptions of these various activation functions are provided in [31]. The three ANN models used to simulate daily GSR series from daily meteorological variables are as follows:

$$\hat{R} = ANN(T_{\max}, T_{\min}, R_a) \tag{5}$$

$$\hat{R} = ANN(H, R_a) \tag{6}$$

$$\hat{R} = ANN(T_{\max}, T_{\min}, H, R_a) \tag{7}$$

where $T_{\max}$ and $T_{\min}$ are daily maximum and minimum temperatures (°K) and H is daily mean relative humidity in a given day. The ANN represents the three-layer feed forward ANN trained by the BRBP algorithm. The R_a is the solar irradiation on a horizontal surface at the top of the atmosphere, which is a function of latitude and Julian day of a site. It is calculated by using the standard geometric method provided by [33]. The details of the method are also available in [20]. The three models are called ANN(T), ANN(H), and ANN(TH) hereafter based on employed input variables. The numbers of hidden nodes selected for the ANN(T), ANN(H), and ANN(TH) are 4, 3, and 5, respectively, based on a trial-and-error procedure.

Daily GSR at ungauged stations that measure other predictor variables can be simulated using a regional ANN-based model. This model is calibrated based on all available GSR and meteorological observations for the region of interest, which allows for the simulation of GSR at ungauged stations where covariables are available. For instance, Fortin et al. [17] and Jeong et al. [20] tested a regional ANN-based model to simulate daily GSR for regional areas located in eastern Canada. They calibrated this model using observations obtained from a set of stations and validated the model using those obtained from a different station set. Regional ANN(T), ANN(H), and ANN(TH) models are also considered in this study using a leave-one-out training procedure. In this approach, the regional ANN models are trained for a given station using observations of all the remaining stations for the calibration period, which is repeated for all stations. In the regional ANN models, mean GSR varies according to R_a, which is a function of the latitude of each station.

2.3 Model evaluation measures

Simulation performances are evaluated using the mean bias error (MBE), root mean square error (RMSE), and R-square (coefficient of determination). The MBE and RMSE are given by the following equations:

$$\text{MBE} = \frac{1}{m}\sum_{i=1}^{m}(\hat{R}_i - R_i) \tag{8}$$

$$\text{RMSE} = [\frac{1}{m}\sum_{i=1}^{m}(\hat{R}_i - R_i)^2]^{0.5} \tag{9}$$

where R_i and $\hat{R}_i$ are observed and simulated daily GSR values and m is the record length. R-square (coefficient of determination) is the squared value of the (Pearson's product-moment) linear correlation coefficient between observed and simulated values. It can provide the proportion of explained variance of observations by an applied model and is defined by the following equation:

$$\mathrm{r}^2 = 1 - \frac{\sum_{i=1}^{m}(\hat{R}_i - R_i)^2}{\sum_{i=1}^{m}(R_i - \overline{R})^2} \tag{10}$$

where $\overline{R}$ is the mean of the observed GSR values.

3. Study area and data

Daily GSR, maximum and minimum temperatures, as well as mean relative humidity are obtained from 13 meteorological stations of Environment Canada (EC) located between latitude 45.1°N to 50.3°N and longitude 64.2°W to 79.0°W (i.e. Southern Quebec, Canada). The daily GSR and the two predictor variables are obtained for the analysis period from 2003 to 2010. Figure 1 shows the locations of the 13 stations across southern Quebec, which have less than 10 % of missing data of daily maximum and minimum temperatures, relative humidity, and GSR for the analysis period. The figure also distinguishes the GSR stations excluded from this analysis due to more than 10 % of missing values of any of the three previously mentioned variables. Stations recording daily maximum and minimum temperatures and relative humidity are presented when they have less than 50 % of missing data for the analysis period. The south of Quebec is the most populated and productive area in the province and has higher density of observation stations than the rest of the province. The three stochastic simulation models are calibrated and validated on the 2003-2007 and 2008-2010 periods, respectively. The three geostatistical interpolation models interpolate the daily GSR for each observation station by using the leave-one-out cross-validation method for the 2008-2010 period. Performances of the six models are finally compared for the 2008-2010 validation period at the 13 selected stations.

Table 1 presents the information (station identification numbers, latitudes, longitudes, and altitudes) of the 13 stations in ascending order of their latitudes. Annual and seasonal (i.e., DJF for winter, MAM for spring, JJA for summer, and SON for autumn) averages of daily GSR for the 2003-2010 period are also provided. In general, it is known that GSR decreases as latitudes increase; however, the annual or seasonal GSR of the stations do not show a clear decrease as their latitudes increase because the study area covers a small range of latitude (5.2 degree). Furthermore, some stations are located in complex climate conditions directly affected by the St-Lawrence River and convections from the Atlantic Ocean (i.e., stations 8, 9, 11, and 13) or from the continent (i.e., station 10). As daily GSR and predictor variables are not linearly correlated, linear correlation coefficients between the solar transmissivity (i.e., the ratio of incoming GSR on the surface of the earth to solar irradiation at the top of the atmosphere) and diurnal temperature range (DTR; T_{max} - T_{min}) $\gamma(R/R_a, \mathrm{DTR})$ series as well as between solar transmissivity and daily mean relative humidity $\gamma(R/R_a, H)$ series are presented. The solar transmissivity and DTR have positive correlations since a cloudy day has smaller GSR, and also a smaller DTR due to a lower T_{max} during the day by blocking sunlight as well as a higher T_{min} during the night by preventing radiative cooling, when compared to a clear day. However, the solar transmissivity and mean daily relative humidity are negatively correlated since a clear day has less humidity than a cloudy day. Correlations between daily GSR and DTR and relative humidity of station 11 are weaker than those of the other stations. This station is located on the south shore of the Lower St-Lawrence valley, which has complex climate conditions affected by the river and convections from the continent and the Atlantic Ocean.

4. Results

4.1 Comparison of model performances

Table 2 presents performance measures of the geostatistical interpolation models for each station for the 2008-2010 validation period. The NN, which is the simplest approach, yields the worst performance, whereas the OK, which is the most sophisticated approach, shows the best performance, although there is a larger magnitude of MBE for OK than for IDW. The three models generally produce larger MBE at stations 10, 12, and 13, which have larger differences in annual mean GSR values compared to the other stations (see Table 1 for values of annual mean GSR and Figure 1 for station locations). The three models yield small RMSEs at stations located in the high density area (i.e., stations 1 to 9), whereas they yield large RMSEs at stations located in the low density area (i.e., stations 10 to 13). In this low density area, the nearest stations to the stations 10-13 are located within a distance of 482.0, 235.2, 236.1, and 363.8 km, respectively, whereas those to the stations 1-9 are located within 100 km.

It is notable that the performance of the geostatistical interpolation models depends on the density of the network of stations and on the statistical homogeneity of GSR values.

Table 3 presents performances of the stochastic simulation models for each station for the calibration and validation periods. The differences of the performances between the calibration and the validation periods are modest for each model and for each station, implying that the three models are calibrated well without overfitting and that they have good generalization ability for a new data set. Average differences between the two periods are 0.35 MJ/m^2/day for MBE, 0.20 MJ/m^2/day for RMSE, and 1.8 % for R-square. Among the three stochastic simulation models, the ANN(TH) uses both temperature and relative humidity as input variables and yields the best performance. The ANN(T) and the ANN(H), which employ either temperature or relative humidity as an input variable, yield similar performances for all stations, except for the station 11, which showed the weakest correlations between daily GSR and predictors among the selected stations (Table 1).

Figure 2 compares RMSEs of the geostatistical interpolation and the stochastic simulation models for each station at annual and seasonal scales for the validation period. The geostatistical interpolation models generally show better performance than the stochastic simulation models for the stations located in the high density area (i.e., stations 1 to 9). However, these models perform differently for the stations located in the low density area (i.e., stations 10 to 13). The poor performances of the geostatistical interpolation models in the low density area are expectable as the models use spatial correlations, which exponentially decrease as distance increase. Especially in spring and summer, RMSEs of the geostatistical models tend to be larger at stations 10, 12, and 13 than the stochastic simulation models, indicating that spatial correlation structures of GSR are weaker in spring and summer than in winter and autumn. However, the stochastic simulation models have similar performances for all stations, except for the station 11, as they only use at-site relationship between the daily GSR and the input variables.

Figure 3 presents scatter plots between observed and simulated daily GSRs for the validation period and for stations 7 and 13, which are located in the high density and the low density (i.e., north-eastern boundary) areas, respectively. In Figures 3a to 3f, the geostatistical interpolation models show better agreement with the 1:1 line than the stochastic simulation models at the station 7. As shown in Tables 2 and 3, the OK model yields the best performance among the six models at this station. However, the geostatistical interpolation models tend to overestimate the observed values at the station because, on average, daily GSRs at the station are smaller than its neighboring stations (see Table 1). In Figures 3g to 3l, the geostatistical interpolation models show worse agreement with the 1:1 line than the stochastic simulation models at the station 13. The ANN(TH) model yields the best performance among the six models.

4.2 Guidelines for model selection

RMSEs and R-squares of daily GSR series between a target and a neighboring station versus their distance for all possible pairs of stations are presented in Figure 4, at an annual and seasonal scales for the 2003-2010 period. In other words, the RMSEs and R-squares of the NN method are calculated, under the assumption that the pair of stations includes the target station and its nearest neighbor. Trend lines of RMSEs and R-squares are estimated by the logarithmic and exponential functions respectively using the non linear least square algorithm. Equations and R-squares of the trend lines for annual and season scales are provided in the figures. Therefore, the trend lines provide approximate RMSEs or R-squares of the NN method for a target station with its nearest neighbor on the study area. For instance, according to the equation presented in Figure 4a, if an observed daily GSR value is available at the nearest neighboring station located at a distance of 200km from a target station, the NN method can approximately simulate the daily GSR at the target station with an expected RMSE of 4.3 MJ/m^2/day at an annual scale. Spatial correlation strengths vary between seasons. For instance, in winter and autumn, the spatial correlation structures are stronger than those in spring and summer. The study area usually shows more homogenized weather and solar radiation conditions in winter and autumn compared to spring and summer seasons because of less convection from Atlantic and/or continental sources.

Using the equations presented in Figure 4, a threshold distance (TD) between a target and its nearest neighboring station can be estimated according to a desired level of performance (i.e., RMSE or R-square) based on the NN model. In Table 4, estimated TDs of the NN model are presented based on the RMSEs of each ANN(T), ANN(H), and ANN(TH) models presented in Table 3. Based on the table, worse performances of the NN models are expected than the stochastic simulation models at stations 10, 12, and 13 as their nearest neighboring stations are located further than their TDs. Similarly, the NN model can yield slightly better performance than ANN(T) and ANN(H), but it can yield a worse performance than ANN(TH) annually at station 11. This can be explained by the NN model requiring nearest neighboring stations to be within 263 km for ANN(T), 245 km for ANN(H), and 212 km for ANN(TH) at an annual scale, but the nearest station (i.e., station 9) is actually at a distance of 235.2 km from station 11.

The TDs presented in Table 4 can be used as a guideline to select an approach between geostatistical interpolation and stochastic simulation models by comparing estimated TDs to the distances of the nearest neighboring stations when filling in missing values and extending record length of daily GSR is required at an observation station. There are three possible cases; (1) TD > distance of nearest neighboring station; (2) TD ≈ distance of nearest neighboring station; (3) TD < distance of nearest neighboring station. For the first case, applying the geostatistical interpolation models is recommended. For instance, on average, better annual performances of the NN model can be expected than the ANN(T), ANN(H), and ANN(TH) when a nearest neighboring station is within 162, 164, and 121 km, respectively. However, the availability of predictor variables (i.e., temperature and/or humidity) of statistical simulation models and the seasonal spatial correlation strengths of geostatisical interpolation models should be considered to select an optimal approach. As ANN(TH) yields better performance than ANN(T) or ANN(H), the former's TD is shorter than the latter's. Shorter TDs are estimated in summer compared to the winter and autumn seasons due to a weak spatial correlation structure in summer. In the second case, applying more sophisticated geostatistical interpolation models (e.g., IDW and OK) than the NN model is recommended. As an example, at station 7, the IDW and OK models yield better performances, whereas the NN model yield a worse performance compared to the ANN(TH) annually (see Figure 3a). Finally, in the third case, applying stochastic simulation models is recommended as they generally can perform better than geostatistical interpolation models. Since the best performance model cannot always be applied for a specific period at a selected station due to a lack of available predictor variables and observed GSR values of neighboring stations, the TD criterion of the proposed guideline can be used to suggest an optimal approach. The guideline and TD can also be used for other GSR stations that were excluded in this analysis due to short record-length (Figure 1).

Under the assumption that a target station has only predictor variables, regional stochastic simulation models are developed using the GSR and predictor variables measured at the other stations. Table 5 presents annual performances of regional models for the 13 stations and their TDs to the nearest neighboring stations to produce similar RMSEs to the regional models. Among the regional ANNs, ANN(TH) yields the best performance while regional ANN(T) and ANN(H) yield similar performances to each other. Again, station 11 shows the worst performance among the 13 stations. The RMSEs of the regional models are 0.21~0.27 MJ/m^2/day larger than those of the at-site models. The worse performances of regional ANNs are reasonable compared to the at-site ANNs as the regional ANNs at each site do not use the observed GSR data of that site for the model calibrations. Consequently, the TDs of the regional models are also 14.9~29.1 km longer than those of the at-site models. These TD values and the ones presented in Table 5 can thus be used to select an appropriate approach between geostatistical interpolation and regional ANN simulation approaches in order to estimate daily GSR at ungauged (or short-record) stations.

5. Concluding remarks

Geostatistical interpolation and stochastic simulation approaches are compared in this study to fill in missing values and to extend short record length of the daily global solar radiation (GSR). However, it is notable that the comparison is only based on the performances of two approaches because they have different application constraints and algorithms to each other. For instance, geostatistical interpolation approaches provide interpolated values at any point in a region including a target station; however, they need observations of daily GSR on the other stations located near the target station to estimate the spatial correlation structure. Stochastic simulation approaches provide estimated values only at the target station using observed daily GSR series as a dependant variable, and daily temperatures as well as humidity series as independent variables.

The simplest nearest neighbor (NN) model yields the worst performance, whereas the most sophisticated ordinary kriging (OK) model shows the best performance among the three geostatistical interpolation models. The three geostatistical interpolation models generally yield smaller RMSEs at stations located in the high density area (i.e., stations 1-9) than those located in the low density area (i.e., stations 10-13). The difference of the performances of geostatistical interpolation models between the high and low density areas can be explained by the exponential decrease of the spatial correlations between stations as the distance increase. Among the three at-site stochastic simulation approach models, the ANN(TH) yields better performance than the ANN(T) and ANN(H), while the ANN(T) and ANN(H) yield similar performances to each other. The three stochastic simulation models produce similar performances for all stations, except for the station 11, which is exposed to a complex climate and showed weaker relationships between GSR and predictors. Regional stochastic simulation models can simulate daily GSR series at stations, where only predictor variables are available; however, the performances of the regionalized models are worse than the at-site models.

In the comparison between the geostatistical interpolation and the stochastic simulation models, the geostatistical models perform better at stations located in the high density area, but they perform worse at stations located in the low density area, compared to the stochastic simulation models. Equations that can approximately estimate the RMSE and R-square based on the NN model using the distance between a target and its nearest neighboring station are presented. By using these equations, a guideline is suggested to select an approach between the geostatistical interpolation and the stochastic simulation approaches. A stochastic simulation approach is recommended when the distance between a target and its nearest neighboring station is longer than the threshold distance (TD) estimated according to the RMSE of a stochastic simulation model. In the opposite case, when the TD is longer than the distance between a target and its nearest neighboring station, a geostatistical interpolation approach is recommended. When the TD is similar to the distance between a target and its nearest neighboring station, more sophisticated geostatistical interpolation models (e.g., IDW and OK) have generally proven to perform better than a stochastic simulation model in this study.

Although, this study suggests a guideline to select an appropriate simulation approach for daily GSR between geostatistical interpolation and stochastic simulation approaches, the guideline is dependent on the spatial correlation strength of daily GSRs and the at-site correlation strength between daily GSRs and the predictor variables. It is proved that spatial correlation strengths for seasonal scales have stronger in winter and autumn compared to those in spring and summer in the study area. Simulation of sub-daily GSR will be considered in future work as it is generally more important than daily GSR to estimate solar energy output due to the non-linear relationship between the radiance and the energy output.

Acknowledgements

This study was financed by a team grant from Ouranos to Yves Gratton, André St-Hilaire, Isabelle Laurion, and J.-C Auclair and by NSERC Discovery grants to André St-Hilaire, Isabelle Laurion, and Yves Gratton. We would like to thank David Huard from Ouranos for his help.

References

[1] H.W. Cutforth, D. Judiesch, Long-term changes to incoming solar energy on the Canadian Prairie, Agricultural and Forest Meteorology 145 (2007) 167-175.

[2] S. Rehman, S.G. Ghori, Spatial estimation of global solar radiation using geostatistics, Renewable Energy 21 (2000) 583-605.

[3] G.G. Merino, D. Jones, D.E. Stooksbury, K.G. Hubbard, Determination of semivariogram models to krige hourly and daily solar irradiance in western Nebrask, Journal of applied meteorology 40 (2001) 1085-1094.

[4] J. Mubiru, K. Karume, M. Majaliwa, E.J.K.B. Banda, T. Otiti, Interpolating methods for solar radiation in Uganda, Theoretical and Applied Climatology 88 (2007) 259-263.

[5] H. Apaydin, F.K. Sonmez, Y.E. Yildirim, Spatial interpolation techniques for climate data in the GAP region in Turkey, Climate Research 28 (2004) 31-40.

[6] H. Alsamamra, J.A. Ruiz-Arias, D. Pozo-Vázquez, J. Tovar-Pescador, A comparative study of ordinary and residual kriging techniques for mapping global solar radiation over southern Spain, Agricultural and Forest Meteorology 149 (2009) 1343-1357.

[7] J.A. Ruiz-Arias, D. Pozo-Vázquez, F.J. Santos-Alamillos, V. Lara-Fanego, J. Tovar-Pescador, A topographic geostatistical approach for mapping monthly mean values of daily global solar radiation: A case study in southern Spain, Agricultural and Forest Meteorology 151 (2011) 1812– 1822.

[8] Zelenka, G. Czeplak, V. D'Agostino, J. Weine, E. Maxwell, R. Perez, Techniques for supplementing solar radiation network data, In Report IEA Task 9, Vol. 2, Report No. IEA-SHCP-9D-1, 1992.

[9] C.L. Goodale, J.D. Aber, S.V. Ollinger, Mapping monthly precipitation, temperature, and solar radiation for Ireland with polynomial regression and a digital elevation model,

Climate Research 10 (1998) 35-49.

[10] D. Leckner, The spectral distribution of solar radiation at the earth's surface-elements of a model, Solar Energy 20 (1978) 143-150.

[11] De La Casiniere, A.I. Bokoye, T. Cabot, Direct solar spectral irradiance measurements and updated simple transmittance models, Journal of Applied Meteorology 36 (1997) 509-520.

[12] Gueymard, A two-band model for the calculation of clear sky solar irradiance, illuminance and photosynthetically active radiation at the earth's surface, Solar Energy 43 (1989) 253-265.

[13] R. Mahmood, K.G. Hubbard, Effect of time of temperature observation and estimation of daily solar radiation for the northern Great Plains, USA, Agronomy Journal 94 (2002) 723-733.

[14] Weiss, C. Hays, Simulation of daily solar irradiance, Agricultural and Forest Meteorology 123 (2004) 187-199.

[15] M. Trnka, Z. Žalud, J. Eitzinger, M. Dubrovský, Global solar radiation in Central European lowlands estimated by various emprical formulare, Agricultural and Forest Meteorology 131 (2005) 54-76.

[16] F.S. Tymvios, C.P. Jacovides, S.C. Michaelides, C. Scouteli, Comparative study of Angstrom's and artificial neural networks' methodologies in estimating global solar radiation, Solar Energy 78 (2005) 752-762.

[17] J.G. Fortin, F. Anctil, L.E. Parent, M.A. Bolinder, Comparison of empirical daily surface incoming solar radiation models, Agricultural and Forest Meteorology 148 (2008) 1332-1340.

[18] M. Benghanem, A. Mellit, S.N. Alamri, ANN-based modelling and estimation of daily

global solar radiation data: A case study, Energy Conversion and Management 50 (2009) 1644-1655.

[19] X. Liu, X. Mei, Y. Li, Q. Wang, J.R. Jensen, Y. Zhang, J.R. Porter, Evaluation of temperature-based global solar radiation models in China, Agricultural and Forest Meteorology 149 (2009) 1433-1446.

[20] D.I. Jeong, A. St-Hilaire, Y. Gratton, C. Bélanger, C. Saad, Simulation and Regionalization of Daily Global Solar Radiation: A Case Study in Quebec, Canada, Atmosphere-Ocean 54 (2016) 117-130.

[21] J. Mubiru, E.J.K.B. Banda, Estimation of monthly average daily global solar irradiation using artificial neural networks, Solar Energy 82 (2008) 181-187.

[22] Y. Jiang, Prediction of monthly mean daily diffuse solar radiation using artificial neural networks and comparison with other empirical models, Energy Policy 36 (2008) 3833-3837.

[23] M.A. Behrang, E. Assareh, A.R. Noghrehabadi, A. Ghanbarzadeh, New sunshine-based models for predicting global solar radiation using PSO (particle swarm optimization) technique, Energy 36 (2011) 3036-3049.

[24] K.L. Bristow, G.S. Campbell, On the relationship between incoming solar radiation and daily maximum and minimum temperature, Agric. For. Meteolo. 31 (1984) 159-166.

[25] R. De Jong, D.W. Stewart, Estimating Global Solar-Radiation from Common Meteorological Observations in Western Canada, Canadian Journal of Plant Science 73 (1993) 509-518.

[26] G.P. Podestá, L. Núñez, C.A. Villanueva, M.A. Skansi, Estimating daily solar radiation in the Argentine Pampas, Agricultural and Forest Meteorology 123 (2004) 41-53.

[27] M. Rivington, G. Bellocchi, K.B. Matthews, K. Buchan, Evaluation of three model

estimations of solar radiation at 24 UK stations, Agricultural and Forest Meteorology 132 (2005) 228-243.

[28] M.G. Abraha, M.G. Savage, Comparison of estimates of daily solar radiation from air tempreature range for application in crop simulations, Agriculatual and Forest Meteorology 148 (2008) 401-416.

[29] E.H. Isaaks, R.M. Srivastava, An introduction to applied geostatistics, Oxford University Press, New York, USA, 1989.

[30] P. Goovaerts, Geostatistics for Natural Resources Evaluation, Oxford University Press, New York, USA, 1997.

[31] S. Haykin, Neural networks. MacMillan College Publishing Company, Hamilton, Ontario, Canada, 1994.

[32] D. Fletcher, E. Goss, Forecasting with neural networks: an application using bankruptcy data, Information and Management 24 (1993) 159-167.

[33] W.K. Sellers, Physical climatology. University of Chicago Press, Chicago, USA, 1965.

Table 1

Station identification number, location information (latitude, longitude, and altitude) as well as annual and seasonal averages of daily GSR of the selected stations for the 2003-2010 analysis period. Linear correlation coefficients between solar transmissivity and diurnal temperature range (DTR) $\gamma(R/R_a, \mathrm{DTR})$ as well as between solar transmissivity and relative humidity $\gamma(R/R_a, H)$ are also provided.

No.	Station #	Lat (°N)	Lon (°W)	Altitude (m)	Average GSR (MJ/m^2/day) Annual	Winter	Spring	Summer	Autumn	$\gamma(\frac{R}{R_a}, \mathrm{DTR})$	$\gamma(\frac{R}{R_a}, H)$
1	7022579	45.05	-72.86	152.4	11.61	4.97	14.58	18.39	8.51	0.53	-0.63
2	702FQLF	45.12	-74.29	49.1	12.83	6.21	16.00	19.91	9.21	0.54	-0.66
3	702LED4	45.29	-73.35	43.8	13.07	6.44	16.49	20.27	9.09	0.50	-0.63
4	7024280	45.37	-71.82	181.0	11.63	5.46	14.53	18.04	8.49	0.58	-0.65
5	702327X	45.72	-73.38	17.9	12.79	6.45	16.10	19.49	9.11	0.52	-0.68
6	7025442	46.23	-72.66	8.0	12.72	6.28	16.07	19.65	8.89	0.52	-0.62
7	7011983	46.69	-71.97	61.0	11.67	5.72	14.99	17.83	8.13	0.63	-0.68
8	701Q004	46.78	-71.29	91.4	11.50	5.44	15.26	17.77	7.54	0.53	-0.72
9	7041JG6	47.08	-70.78	6.0	11.97	5.56	15.29	18.57	8.48	0.51	-0.65
10	7086716	48.25	-79.03	318.0	11.83	5.39	15.96	18.49	7.49	0.54	-0.73
11	7056068	48.51	-68.47	4.9	12.16	4.99	16.02	19.49	8.15	0.29	-0.47
12	7065639	48.84	-72.55	137.2	12.63	6.21	17.09	18.78	8.44	0.52	-0.68
13	7044328	50.27	-64.23	11.0	11.28	4.18	15.46	17.92	7.56	0.58	-0.59
Avg.		46.86	-72.05	83.2	12.13	5.64	15.68	18.82	8.39	0.52	-0.65

Table 2

Performance measures of the three geostatistical interpolation models for the 2008-2010 period.

station #	MBE (MJ/m^2/day)			RMSE (MJ/m^2/day)			R-square (100^{-1}%)		
	NN	IDW	OK	NN	IDW	OK	NN	IDW	OK
1	-0.67	-0.33	-0.45	2.53	2.34	2.15	0.91	0.92	0.94
2	0.08	0.51	0.16	2.24	2.53	2.38	0.92	0.90	0.91
3	0.63	0.40	0.15	2.44	2.03	2.17	0.92	0.94	0.93
4	-0.26	-0.50	-0.37	2.34	2.60	2.50	0.92	0.89	0.90
5	-0.24	0.18	-0.14	2.36	1.80	1.86	0.91	0.94	0.94
6	1.46	0.61	0.71	2.99	2.01	1.98	0.89	0.94	0.95
7	-1.03	-1.24	-1.14	2.32	2.17	2.01	0.93	0.94	0.95
8	-0.37	-0.13	0.09	2.24	1.91	1.89	0.92	0.95	0.94
9	0.26	0.22	0.26	2.60	2.68	2.49	0.90	0.90	0.91
10	-0.78	-0.42	-0.48	5.50	5.49	5.10	0.58	0.55	0.60
11	-0.48	-0.32	-0.20	4.55	4.18	3.79	0.73	0.76	0.80
12	0.34	0.55	0.67	4.93	4.22	3.75	0.67	0.73	0.79
13	-0.45	-0.82	-0.66	5.51	5.80	5.33	0.62	0.55	0.61
avg.	-0.12	-0.10	-0.11	3.27	3.06	2.88	0.83	0.84	0.86

Table 3

Performance measures of the three stochastic simulation models during the 2003-2007 calibration and the 2008-2010 validation periods.

	MBE ($MJ/m^2/day$)			RMSE ($MJ/m^2/day$)			R-square (100^{-1}%)		
	ANN(T)	ANN(H)	ANN(TH)	ANN(T)	ANN(H)	ANN(TH)	ANN(T)	ANN(H)	ANN(TH)
Calibration period (2003~2007)									
1	0.03	0.31	0.02	3.76	3.82	3.19	0.76	0.76	0.83
2	-0.04	0.05	0.00	3.94	3.88	3.18	0.78	0.78	0.86
3	-0.06	-0.04	-0.04	4.48	4.49	3.86	0.74	0.74	0.81
4	-0.03	0.05	0.01	3.77	3.77	3.06	0.75	0.75	0.84
5	0.03	0.05	-0.01	3.77	3.72	3.14	0.79	0.79	0.85
6	0.00	-0.07	0.07	3.88	4.12	3.26	0.77	0.74	0.84
7	0.09	0.02	-0.04	3.77	3.94	3.01	0.78	0.76	0.86
8	-0.04	0.00	0.00	4.02	3.62	3.16	0.75	0.80	0.85
9	0.00	-0.01	0.01	4.39	4.35	3.64	0.72	0.72	0.80
10	-0.01	0.03	-0.02	3.62	3.14	2.78	0.80	0.85	0.88
11	0.04	0.00	-0.04	4.87	4.53	4.35	0.68	0.72	0.75
12	-0.07	0.10	0.03	3.83	3.66	3.20	0.78	0.80	0.85
13	-0.10	0.03	0.00	3.75	3.76	3.37	0.80	0.80	0.84
avg.	-0.01	0.04	0.00	3.99	3.91	3.32	0.76	0.77	0.83
validation period (2008~2010)									
1	0.66	0.78	0.53	3.92	4.08	3.37	0.79	0.77	0.84
2	-0.36	-0.09	-0.28	3.89	3.84	3.25	0.77	0.77	0.84
3	-0.75	-0.56	-0.66	3.65	3.80	3.22	0.81	0.79	0.85
4	-0.01	-0.27	-0.21	3.62	3.81	3.03	0.79	0.76	0.85
5	-0.34	0.21	-0.08	3.40	3.36	2.94	0.80	0.81	0.85
6	-0.07	-0.32	-0.09	3.81	3.89	3.12	0.77	0.77	0.85
7	-0.86	-0.42	-0.67	3.42	3.63	2.84	0.81	0.77	0.87
8	0.40	-0.40	-0.06	3.76	3.75	3.28	0.79	0.78	0.83
9	0.26	0.29	0.28	4.24	4.19	3.58	0.74	0.74	0.81
10	0.33	0.50	0.49	3.81	3.46	3.08	0.76	0.81	0.85
11	-0.52	-0.53	-0.60	4.76	4.63	4.32	0.69	0.71	0.75
12	0.36	0.08	0.25	3.77	3.57	3.18	0.78	0.79	0.84
13	0.18	0.70	0.51	3.79	4.11	3.54	0.79	0.77	0.82
avg.	-0.05	0.00	-0.05	3.83	3.85	3.29	0.78	0.77	0.83

Table 4

Threshold distances (TDs; in km) between the target and nearest neighboring station for the NN model to produce same RMSEs as the ANN(T), ANN(H), and ANN(TH), respectively. The TDs are calculated by equations presented in Figure 4 with the RMSEs of the three stochastic simulation models at each station and each time scale for the validation period presented in Table 3.

	ANN(T)					ANN(H)					ANN(TH)				
	annual	winter	spring	summer	autumn	annual	winter	spring	summer	autumn	annual	winter	spring	summer	autumn
1	167.0	356.1	209.9	109.5	217.7	181.9	212.4	199.2	145.8	249.9	123.7	194.3	146.3	88.5	168.2
2	164.0	303.6	196.6	127.9	183.9	160.1	185.4	224.0	104.7	214.1	115.8	158.6	158.8	75.2	154.6
3	144.0	231.5	176.4	101.1	207.8	156.2	157.3	221.6	94.4	248.4	114.0	121.0	163.2	66.3	183.8
4	142.0	205.5	170.9	108.8	161.6	156.9	130.0	201.3	112.1	211.1	103.0	120.7	133.3	72.1	129.2
5	125.6	155.1	142.1	102.6	154.7	123.4	126.4	145.0	94.7	178.6	98.2	99.2	126.1	71.8	118.8
6	157.2	183.8	197.3	116.7	184.5	163.9	106.6	183.2	134.3	250.4	108.2	116.5	127.8	85.4	135.6
7	126.9	208.6	145.5	84.5	128.6	142.2	119.7	170.5	97.1	132.7	92.7	99.3	120.4	59.3	80.2
8	153.2	260.3	195.0	115.8	168.3	152.1	137.2	207.1	122.2	143.4	117.8	137.5	155.5	96.2	102.8
9	198.4	349.5	231.0	148.1	274.5	192.8	231.4	200.0	186.0	190.6	138.8	207.0	153.5	123.2	138.4
10	157.3	78.1	212.1	113.0	201.8	129.9	112.6	151.9	101.1	184.1	106.1	64.2	135.2	81.0	135.6
11	263.0	162.5	242.8	270.5	313.2	245.6	125.1	253.6	261.4	237.8	206.8	116.2	212.3	216.8	212.2
12	153.8	145.4	174.3	114.8	194.1	137.9	110.0	158.1	109.0	187.5	111.6	107.5	132.1	89.0	124.1
13	155.2	99.4	172.2	153.5	142.3	184.7	102.2	162.6	218.4	134.1	136.1	79.7	138.4	145.7	102.5
avg.	162.1	210.7	189.7	128.2	194.9	163.7	142.8	190.6	137.0	197.1	121.0	124.7	146.4	97.7	137.4

Table 5

Annual RMSEs of regional ANN(T), ANN(H), and ANN(TH) and their threshold distances (TDs) of the nearest stations for the NN model to produce same RMSEs as the regional models. The TDs are calculated by equations presented in Figure 4 with the RMSEs of the three regional models at each station during the validation period.

	regional ANN(T)		regional ANN(H)		regional ANN(TH)	
	RMSE ($MJ/m^2/day$)	TD (km)	RMSE ($MJ/m^2/day$)	TD (km)	RMSE ($MJ/m^2/day$)	TD (km)
1	3.89	164.0	4.14	187.7	3.48	131.4
2	3.97	171.1	3.91	165.8	3.30	119.4
3	3.65	144.3	3.77	153.9	3.19	112.4
4	3.98	172.2	3.93	168.1	3.38	124.6
5	3.46	130.3	3.37	124.0	2.97	99.6
6	3.89	164.4	3.84	160.1	3.16	110.5
7	4.47	225.1	3.64	143.7	3.39	124.8
8	3.75	152.1	4.10	183.4	3.23	114.8
9	4.28	202.7	4.19	193.0	3.63	142.3
10	3.81	157.4	3.72	149.9	3.20	112.7
11	5.66	428.8	5.43	377.4	5.63	421.5
12	3.97	171.2	3.82	158.3	3.44	128.7
13	3.78	154.6	4.92	286.3	4.34	209.2
avg.	4.04	187.6	4.06	188.6	3.56	150.1

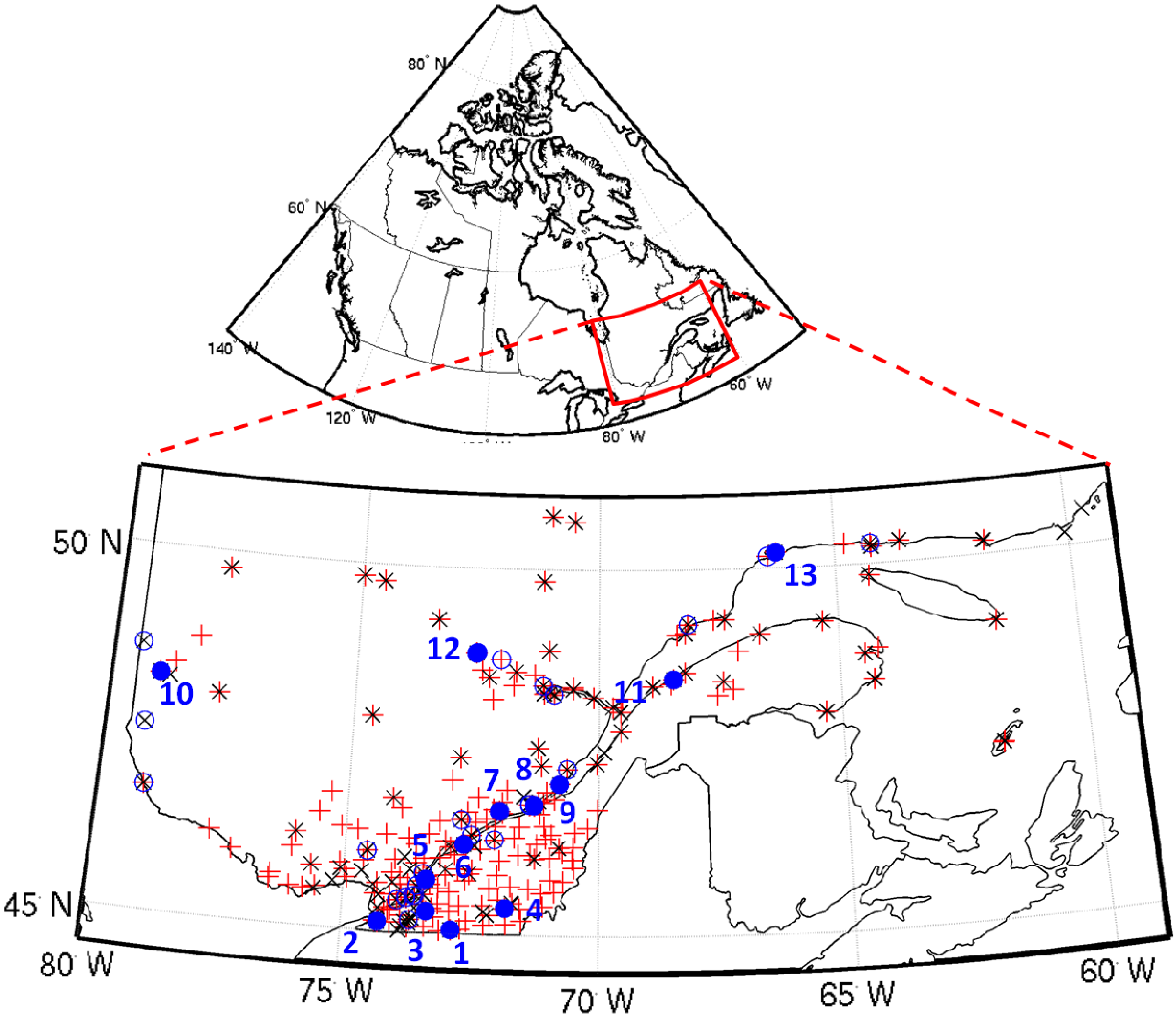


Fig. 1. Map of southern Québec, Canada. Stations are presented by red '+' and black '×' when they have observed daily temperature and relative humidity respectively, when they have less than 50 % of missing data for the common analysis period (from 2003 to 2010). Blue filled circles represent the selected meteorological stations, which have less than 10 % of missing data of daily temperature, relative humidity, and GSR for the common analysis period. Blue open circles represent the GSR stations excluded in this analysis due to more than 10 % missing values of any of the three previously mentioned variables.

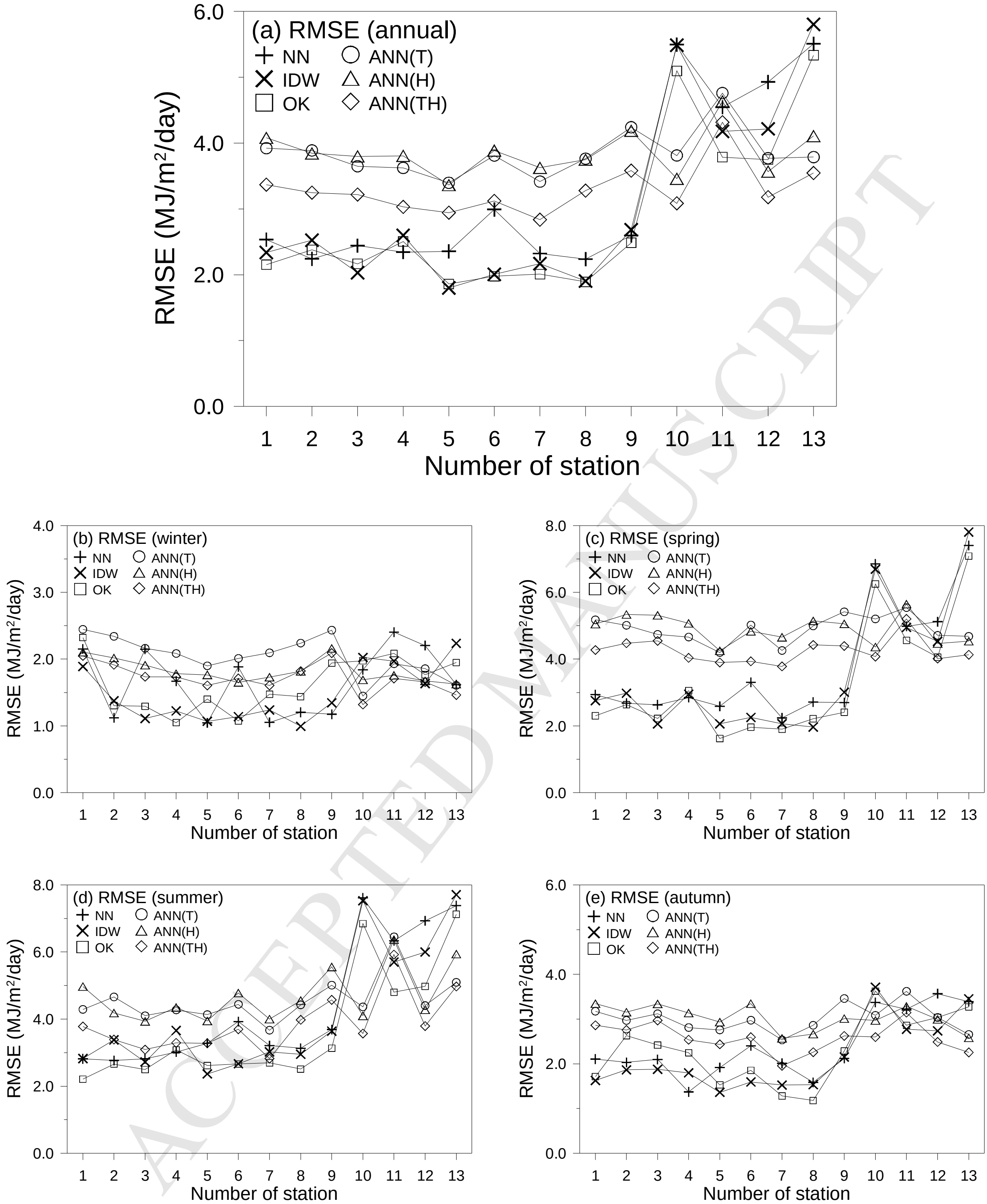


Fig. 2. RMSEs of the three geostatistical interpolation (NN, IDW, and OK) and the three stochastic simulation (ANN(T), ANN(H), and ANN(TH)) models for each stations at (a) annual and (b-e) seasonal scales during the 2008-2010 validation period.

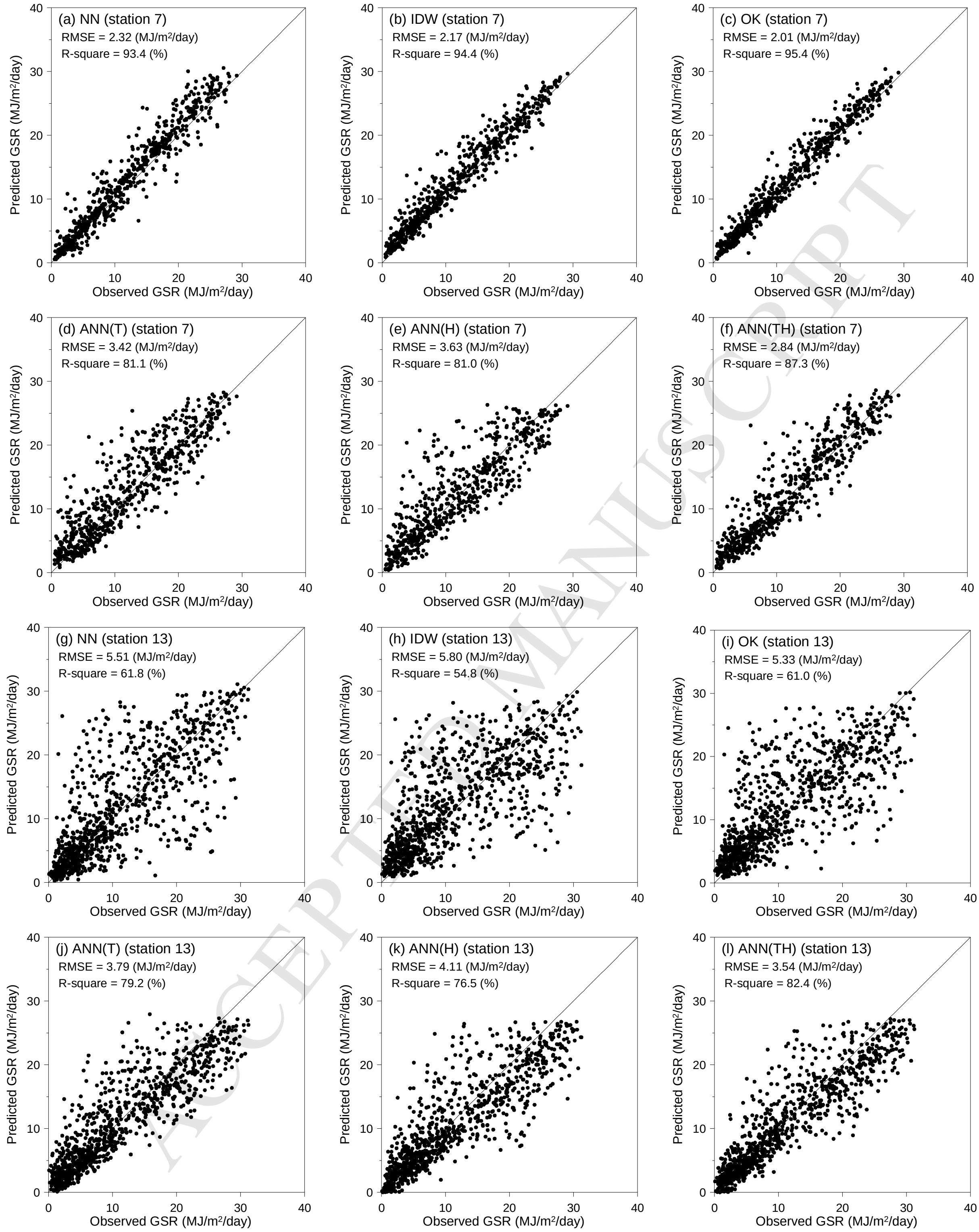
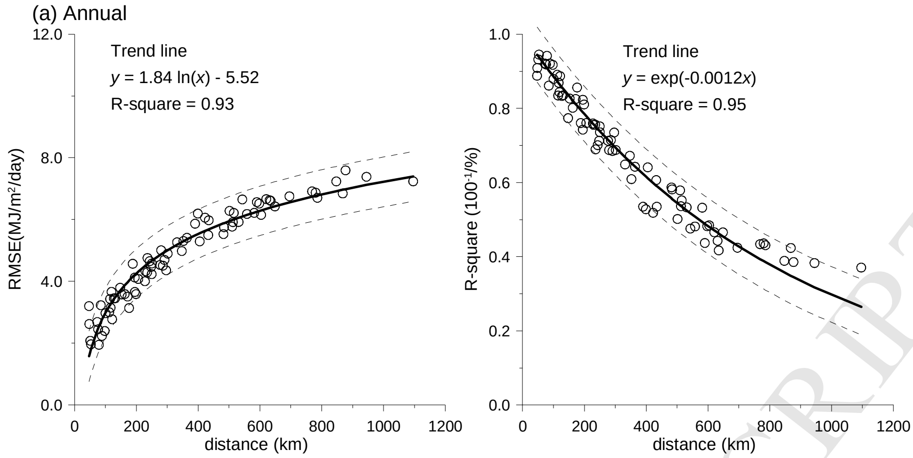
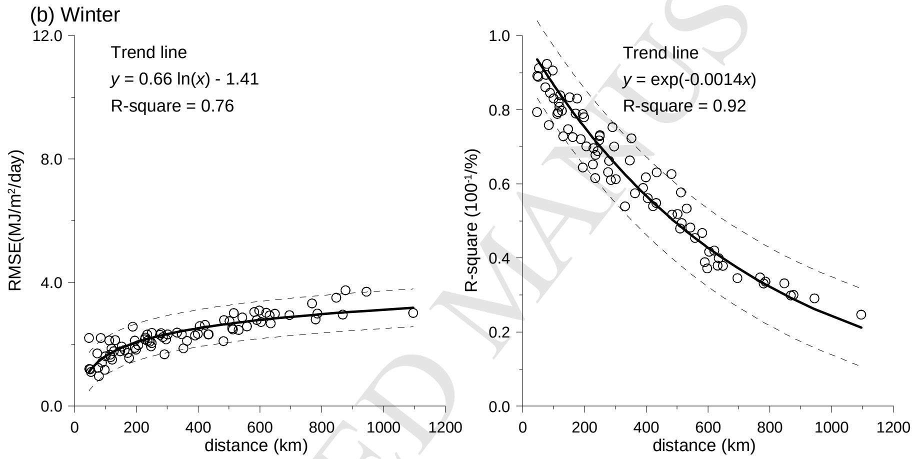
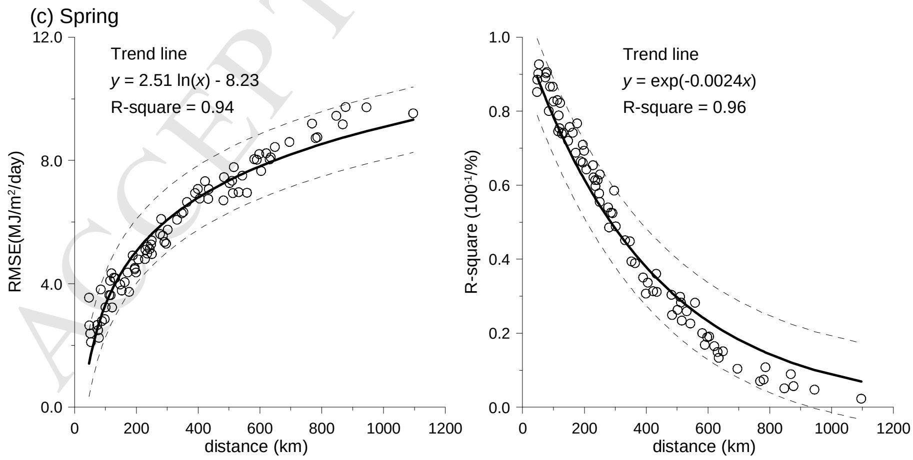


Fig. 3. Scatter plots of daily GSRs between observation and predictions by the three geostatistical interpolation and the three stochastic simulation models at the stations 7 (a-f) and 13 (g-l).


(a) Annual
Trend line
y = 1.84 ln(x) - 5.52
R-square = 0.93
Trend line
y = exp(-0.0012x)
R-square = 0.95
RMSE(MJ/m²/day)
R-square (100⁻¹/%)
distance (km)


(b) Winter
Trend line
y = 0.66 ln(x) - 1.41
R-square = 0.76
Trend line
y = exp(-0.0014x)
R-square = 0.92
RMSE(MJ/m²/day)
R-square (100⁻¹/%)
distance (km)


(c) Spring
Trend line
y = 2.51 ln(x) - 8.23
R-square = 0.94
Trend line
y = exp(-0.0024x)
R-square = 0.96
RMSE(MJ/m²/day)
R-square (100⁻¹/%)
distance (km)

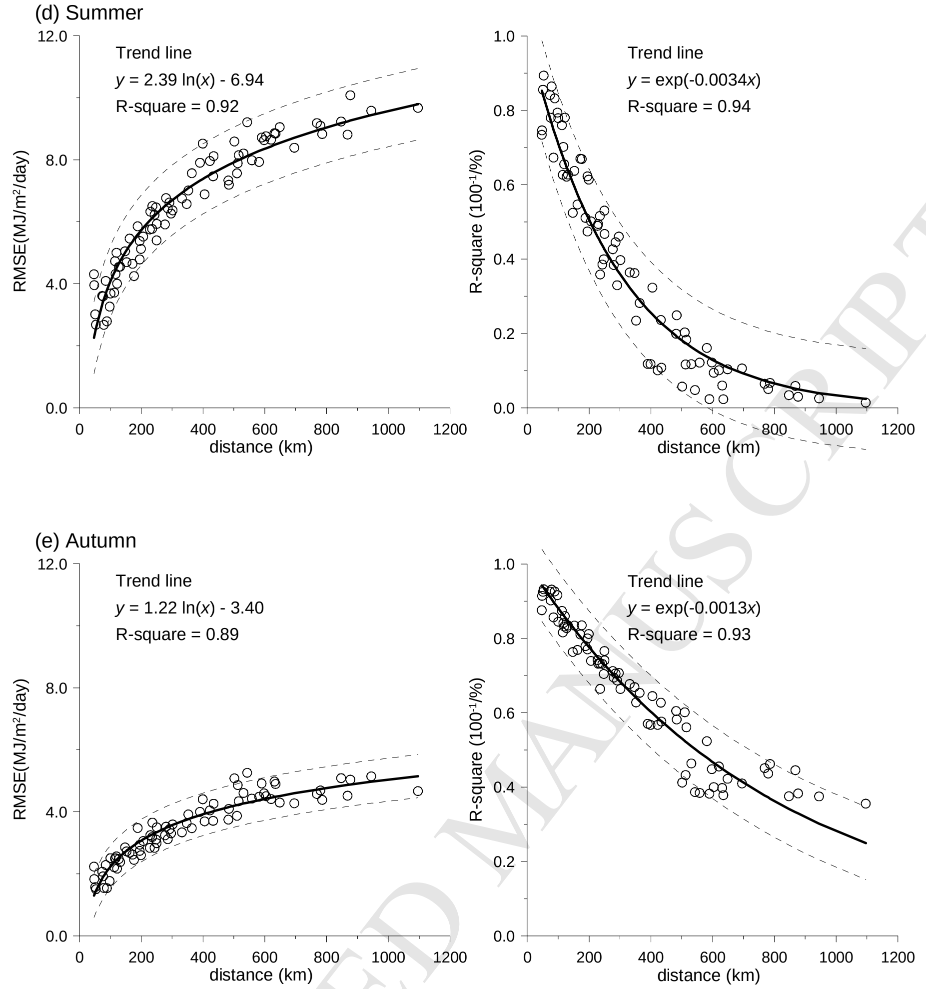


Fig. 4. RMSE and R-square of daily GSR series between a target and its neighboring stations versus the distance between the two stations for all possible combinations during the analysis period from 2003 to 2010. Trend lines of RMSEs and R-squares are estimated by logarithmic and exponential functions, respectively. Equations and R-squares of the trend lines are presented on the figures. The dotted lines represent the 95 % confidence interval of the trend lines.

- Models for estimating daily global solar radiation are investigated.
- Geostatistical interpolation and stochastic simulation approaches are compared.
- Geostatistical models yield better performance at a high density measurement area.
- Stochastic models show better performance at a low density measurement area.
- A guideline to select an optimal estimation approach is then suggested.