Optimal weighting scheme for averaging regional temperature (I)

—— Theoretical analysis

Shen, S.S.P.,

(Department of Mathematical Sciences, University of Alberta, Edmonton, Canada T6G 2G1)

WANG Xiaochun (王晓春), LIANG Youlin (梁幼林)
and LI Rongfeng (李荣凤)

(Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100080, China)

Received February 25, 1995

Keywords: optimal weighting scheme, regional average temperature, arithmetic average.

Inadequacy of a spatial data sampling scheme often causes confusion when considering global, hemispherical and regional temperature averages. Jones *et al.*[1, 2], Hansen *et al.*[3], Vinnikov *et al.*[4] used different methods in their researches for the long-term trend of global or hemispherical surface air temperatures. Optimal weighting scheme was used by Vinnikov *et al.* The weight for each station at different latitude was obtained in the sense of minimum mean square error of average temperature. The weights were then used to compute the global or hemispherical average temperature. Recently, using empirical orthogonal functions, Shen *et al.*[5] showed that the global average temperature can be computed with quite satisfactory accuracy by using only above 60 stations distributed properly over the globe.

Our research will investigate the influence of inadequate spatial sampling on the computation of regional average temperature. The optimal weighting scheme is given in sec. 2. Some special cases of the scheme are analyzed in sec. 3; secs. 4 and 5 show some results of test computations and summarize our conclusions. More comprehensive computational results about the comparison between the optimal weighting scheme and the ordinary arithmetic average will be given in another article.

1 The theory of optimal weighting scheme

Let $\boldsymbol{r}$ be the position of point of question in region C, $\Theta(\boldsymbol{r}, t)$ be the temperature at $\boldsymbol{r}$ and time t, $\Theta_\tau(\boldsymbol{r}, t)$ be the τ-length average of temperature centered at t and at the location $\boldsymbol{r}$, i.e.

$$\Theta_\tau(\boldsymbol{r}, t) = \frac{1}{\tau}\int_{t-\frac{\tau}{2}}^{t+\frac{\tau}{2}} \Theta(\boldsymbol{r}, t')\mathrm{d}t', \tag{1}$$

where τ can be one year, one month, and so on.

The real regional average of temperature averaged over τ-length time in region C is

$$\Theta(t)=\frac{1}{A}\int_C \mathrm{d}\Omega\Theta_\tau(\boldsymbol{r},t), \tag{2}$$

where A is the area of region C, $\mathrm{d}\Omega$ is the integration element. And it should be noted that the above regional average is relevant to τ.

The estimation of $\Theta(t)$ by using observation of n stations is

$$\hat{\Theta}(t)=\sum_{i=1}^{n} w_i\Theta_\tau(\boldsymbol{r}_i,t), \tag{3}$$

or

$$\hat{\Theta}(t)=\frac{1}{A}\int_C \mathrm{d}\Omega W\Theta_\tau(\boldsymbol{r},t), \tag{4}$$

where

$$W=A\sum_{i=1}^{n} w_i\delta(\boldsymbol{r}-\boldsymbol{r}_i), \tag{5}$$

$\boldsymbol{r}_i$ is the position of ith station, $\Theta_\tau(\boldsymbol{r},t)$ the τ-length average of temperature for ith station station centered at t, w_i the weight for ith station, and $\delta(\boldsymbol{r}-\boldsymbol{r}_i)$ the δ function. The following constraint should be imposed on w_i

$$\sum_{i=1}^{n} w_i=1. \tag{6}$$

The mean square error of the estimator $\hat{\Theta}(t)$ for $\Theta(t)$ is

$$\varepsilon^2=\langle(\hat{\Theta}(t)-\Theta(t))^2\rangle, \tag{7}$$

where $\langle\cdot\rangle$ is the ensemble average. Expansion of the above formula leads to

$$\varepsilon^2=\frac{1}{A}\int_C \mathrm{d}\Omega'\int_C \mathrm{d}\Omega'\rho(\boldsymbol{r},\boldsymbol{r}')-\frac{2}{A}\sum_{i=1}^{n} w_i\int_C \mathrm{d}\Omega\rho(\boldsymbol{r},\boldsymbol{r}_i)+\sum_{i,j=1}^{n} w_iw_j\rho(\boldsymbol{r}_i,\boldsymbol{r}_j), \tag{8}$$

where

$$\rho(\boldsymbol{r}_i,\boldsymbol{r}_j)=\langle\Theta_\tau(\boldsymbol{r}_i,t)\Theta_\tau(\boldsymbol{r}_j,t)\rangle, \tag{9}$$

in which we take $\Theta_\tau(\boldsymbol{r}_i,t)$ as stationary.

We try to find the weight for every station by minimizing ε^2 with (6) as a constraint. The weight obtained by this procedure is called the optimal weight. The method of Lagrange multiplier is used with -2Λ as the multiplier for the convenience in the following derivation. We denote

$$F=\varepsilon^2-2\Lambda\left(\sum_{i=1}^{n} w_i-1\right). \tag{10}$$

Using the same method as that in Shen *et al.*, Vinnikov *et al.*, from

$$\frac{\partial F}{\partial w_i}=0 \text{ and } \frac{\partial F}{\partial\Lambda}=0, \tag{11}$$

the optimal weights are the solution of the following linear equations

$$\begin{cases} \sum_{j=1}^{n} w_j \rho(\boldsymbol{r}_i, \boldsymbol{r}_j) - \Lambda = \overline{\rho}(\boldsymbol{r}_i) \quad i = 1, \cdots, n, \\ \sum_{i=1}^{n} w_i = 1, \end{cases} \tag{12}$$

where

$$\overline{\rho}(\boldsymbol{r}_i) = \frac{1}{A} \int_C \mathrm{d}\Omega \rho(\boldsymbol{r}, \boldsymbol{r}_i). \tag{13}$$

If we denote the inverse of the matrix $(\rho_{ij} = \rho(\boldsymbol{r}_i, \boldsymbol{r}_j))$ as (b_{ij}), then the solution for (12) can be expressed as

$$\begin{cases} w_i = \sum_{j=1}^{n} b_{ij} (\Lambda + \overline{\rho}(\boldsymbol{r}_j)), \\ \Lambda = \dfrac{1 - \sum_{i=1}^{n} \sum_{j=1}^{n} b_{ij} \overline{\rho}(\boldsymbol{r}_j)}{\sum_{i=1}^{n} \sum_{j=1}^{n} b_{ij}}. \end{cases} \tag{14}$$

If the temperature data are standardized before the computation, $\rho(\boldsymbol{r}_i, \boldsymbol{r}_j)$ is the correlation coefficient of stations i and j. So the quantity $\overline{\rho}(\boldsymbol{r}_i)$ basically measures the importance of station i when its observation is used to compute the regional average. Considering what we try to compute is the optimal average temperature over region C, and formula for $\overline{\rho}(\boldsymbol{r}_i)$ is also a regional integration, we may conclude that the original problem is changed into a new one in the same form.

2 Some special cases of the optimal weighting scheme

The above theoretical consideration has shown a method to get regional average temperature by the optimal weighting scheme. But if $\overline{\rho}(\boldsymbol{r}_i)$ cannot be computed accurately, then the satisfactory regional average cannot be obtained. The following special cases of optimal weighting scheme will show the relationship of the optimal weighting scheme to the computation method of $\overline{\rho}(\boldsymbol{r}_i)$.

2.1 Using real correlation coefficient to compute $\overline{\rho}(\boldsymbol{r}_i)$

If the correlation coefficients among n stations are used to directly compute $\overline{\rho}(\boldsymbol{r}_i)$, then from the discrete form of (13), we get

$$\overline{\rho}(\boldsymbol{r}_i) = \frac{1}{A} \int_C \mathrm{d}\Omega \rho(\boldsymbol{r}, \boldsymbol{r}_i) = \sum_{j=1}^{n} \frac{\Delta A_j}{A} \rho(\boldsymbol{r}_j, \boldsymbol{r}_i), \tag{15}$$

where ΔA_j is the area of the subregion that can be represented by observation from jth station. From (12), the solution for optimal weights is

$$\begin{cases} w_i = \dfrac{\Delta A_i}{A}, \\ \Lambda = 0. \end{cases} \tag{16}$$

This means that the weight for each station equals the ratio of the area represented by the station to the total area of the region. Specifically, if the area represented by each station is equal to each other, then the solution is

$$\begin{cases} w_i = \dfrac{1}{n}, \\ \Lambda = 0. \end{cases} \tag{17}$$

This means the optimal weighting scheme is debased to the ordinary arithmetic average.

2.2 Using the presumed correlation coefficient field to compute $\overline{\rho}(\boldsymbol{r}_i)$

If a theoretical distribution of correlation coefficient field is presumed, the result of $\overline{\rho}(\boldsymbol{r}_i)$ may be better. Vinnikov *et al.* presumed an empirical correlation coefficient formula for every latitude band of 30 degrees. For 30°—60° N, the presumed formula is

$$\rho(s) = \exp(-0.21 s^{0.893}) J_0(0.852 s), \tag{18}$$

where ρ is the correlation coefficient, s is the distance between stations with 10^3 km as the unit, and J_0 is the Bessel function. Their method renders satisfactory result for computing global average temperature. Since the correlation coefficient of temperature varies with respect to geographic position and time scale[6], their method is clearly not universally applicable.

Our results show, in the case of the computation of regional average, that the relationship of correlation coefficient with distance can be more accurate if expressed by an exponential decrease with respect to the distance square. In this research, the following formula is presumed for the correlation coefficient with ith station as the base point

$$\rho(\boldsymbol{r}, \boldsymbol{r}_i) = a_i \exp\left(-\frac{|\boldsymbol{r}-\boldsymbol{r}_i|^2}{d_i^2}\right), \tag{19}$$

where a_i and d_i^2 are the coefficients relevant to station i and time scale τ.

According to the above assumption, the coefficients a_i and d_i^2 are computed using the real correlation coefficient field. Then $\overline{\rho}(\boldsymbol{r}_i)$ is computed in the designated region. Hence the optimal weight can be obtained by formula (14) and at last the optimal regional average temperature can be computed.

Compared with the ordinary arithmetic average, the optimal average scheme is more complicated. But if we have considered that the coefficients a_i and d_i^2 can be computed quite accurately with sufficient latest observation data, and if the coefficients can be used to compute the regional average when the observation is very sparse and rare, we may conclude that it deserves to compute the optimal weight this way. And it should also be noted that the coefficients are relevant to time scale τ.

3 Computation results

The observation of 23 stations located in the Northeast China for the period of

1961—1990 are used in our test computation. First, the seasonal change is removed by using the original temperature data to subtract the monthly mean over a long term. Then the anomalies are standardized, and correlation coefficient matrix is computed. The least-square-fit method is used to get coefficients a_i and d_i^2 for each station, and then the weight for each station is computed. The optimal regional average temperature anomaly (OTM) is compared with the arithmetic regional average temperature anomaly (TM), where the anomaly means the standardized anomaly.

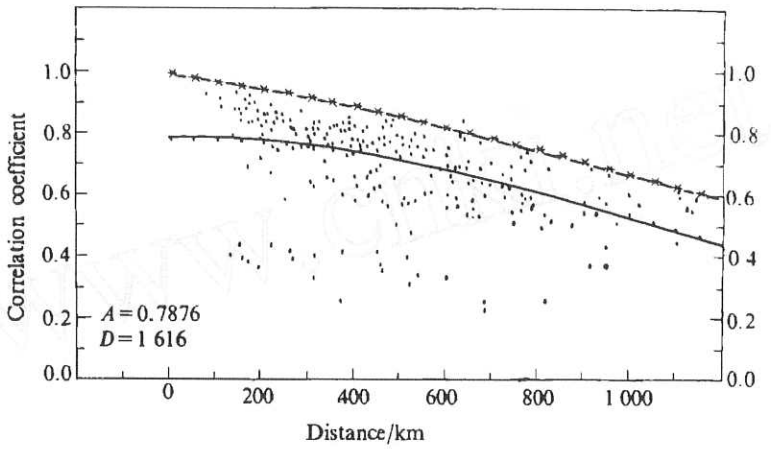

Fig. 1. The relationship of correlation coefficients to distance. —●— $A\exp\left(-\frac{r^2}{D}\right)\left(A=\frac{1}{n}\sum_{j=1}^{n}a_j,\ D=\frac{1}{n}\sum_{j=1}^{n}d_j^2,\right.$ r is the distance between stations$\Big)$, —×— $\exp(-0.21s^{0.893})J_0(0.852s)$ (s is the distance between stations, J_0 is the Bessel function).

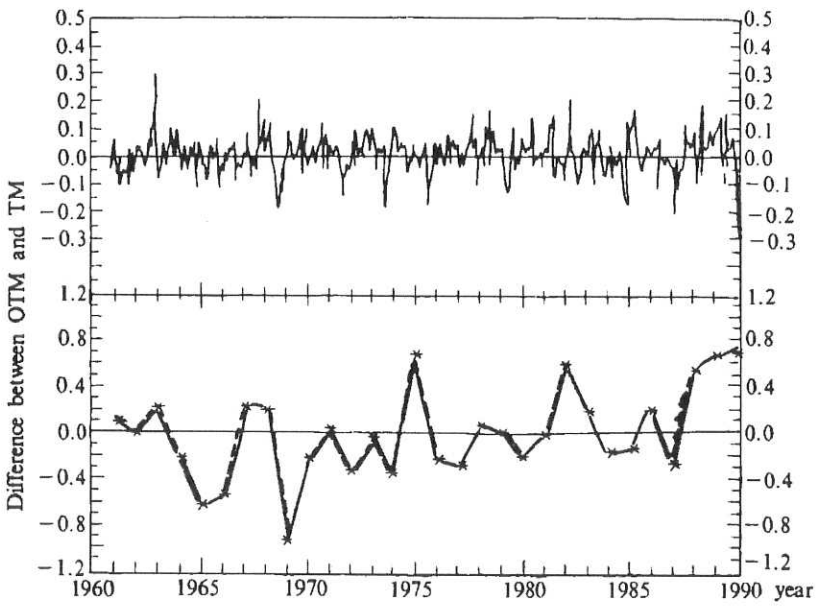

Fig. 2. The annual change of the optimal regional average temperature anomaly (dashed line in lower part) and arithmetic regional average temperature anomaly (solid line in lower part), and the monthly change of their differences (upper part).

The relationship between correlation coefficient and distance is shown in fig. 1. It is clear that, for regional temperature field, our expression of the correlation coefficient is closer to reality than the formula designated by Vinnikov *et al.*

For better expression of the figure, annual changes of the OTM and TM are shown in the lower part of fig. 2. The monthly changes of their differences are shown in the upper part of fig. 2. It can be noted from fig. 2 that their difference is rather small. In fact, their root mean square difference

$$\mathrm{RMSD}=\left(\frac{1}{m}\sum_{i=1}^{n}(\mathrm{OTM}_i-\mathrm{TM}_i)^2\right)^{1/2} \tag{20}$$

is 9.0 percent of the standard deviation of TM. In the above formula, m is the total number of months, OTM_i and TM_i are the OTM and TM for i month. From comparison, we can conclude that the method in our research can get a reasonable regional average.

4 Conclusions

The optimal weighting scheme for the computation of the regional average temperature is derived. The analyses of some special cases of the optimal weighting scheme show that the computation method of $\overline{\rho}(r_i)$ is the crucial part of the scheme.

The test computation, using the observation of 23 stations from Northeast China, shows that the optimal weighting scheme can get a reasonable regional average. The full comparison of the optimal weighting scheme and the arithmetic regional average will be given in another report.

Acknowledgement Thanks are due to Dr Hu Zengzhen, whose suggestions have helped improve the manuscript.

References

1 Jones, P. D., Raper, S. C. B., Bardley, R. S. *et al.*, Northern Hemisphere surface air temperature variations: 1851—1984. *J. Clim. Appl. Meteorol.*, 1986, 25(2): 161.
2 Jones, P. D., Raper, S. C. B., Bradley, R. S. *et al.*, Southern Hemisphere surface air temperature variations: 1851—1984. *J. Clim. Appl. Meteorol.*, 1986, 25(9): 1213.
3 Hansen, J., Lebedeff, S., Global trends of measured surface air temperature. *J. Geophys. Res.*, 1987, 92(D11): 13345.
4 Vinnikov, K. Y., Groisman, P. Y., Lugina, K. M., Empirical data on contemporary global climate changes (temperature and precipitation). *J. Climate.*, 1990, 3(6): 662.
5 Shen, S. S. P., North, G. R., Kim, K. Y., Spectral approach to optimal estimation of the global average temperature. *J. Climate*, 1994 (in press).
6 Kim, K. Y., North, G. R., Surface temperature fluctuations in a stochastic climate model. *J. Geophys. Res.*, 1991, 96(D10): 18573.