Environmental Modelling & Software 66 (2015) 69–77

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

A geostatistics-assisted approach to the deterministic approximation of climate data

Maria Lanfredi [a, *], Rosa Coppola [a], Mariagrazia D'Emilio [a], Vito Imbrenda [a], Maria Macchiato [b], Tiziana Simoniello [a]

[a] Institute of Methodologies for Environmental Analysis, Italian Research Council, C.da S. Loja, Tito, PZ, I-85050, Italy
[b] Department of Physics, "Università Federico II", via Cinthia, Napoli, I-80126, Italy

ARTICLE INFO

Article history:
Received 12 August 2014
Received in revised form
7 December 2014
Accepted 11 December 2014
Available online

Keywords:
Geostatistics
Geographical model
Climatic surface

ABSTRACT

We propose a nonconventional application of variogram analysis to support climate data modelling with analytical functions. This geostatistical technique is applied in the theoretical domain defined by each model variable to detect the systematic behaviours buried in the fluctuations determined by other driving factors and to verify the ability of candidate fits to remove correlations from the data. The climatic average of the atmospheric temperature measured at 387 European meteorological stations has been analysed as a function of geographical parameters by a step-wise procedure. Our final model accounts for non-linearity in latitude with a local-scale residual correlation that decays in approximately ten kilometres. The variance of the residuals from the fitted model (approximately 3% of the total) is mostly determined by local heterogeneity in transitional climates and by urban islands. Our approach is user-friendly, and the support of statistical inference makes the modelling self-consistent.

1. Introduction

Recent satellite remote sensing technologies for Earth Observation (EO) have supplied a large amount of spatial data that are promising for improving our understanding of the climate system. Contextually, the sparse and uneven data provided by ground stations are still an essential source of information on many key variables characterizing climate dynamics. Currently, the collection of data obtained from meteorological networks, which are generally regarded as valid for spatial inferences of the state of the low atmosphere (Geiger et al., 2003), are also used within climatic studies at the planetary scale. As an example, the series of global datasets, HadCRUT, gridded on a 5° × 5° latitude–longitude box grid, has been widely exploited for the evaluation and attribution of climate change (e.g., Brohan et al., 2006; Jones and Stott, 2011; Jones et al., 2012).

Multi-resolution, both in time and in space, provides the standard hierarchical framework for studying the dynamics of the climate system. Because details at different resolutions generally characterize different physical structures, a coarse-to-fine descriptive strategy is used to separate the broad scale context that is properly climatic from the local contexts of weather dynamics.

In regional studies, numerical models (Rummukainen, 2010; Feser et al., 2011) add details to global-scale climate models, thus improving simulations and forecasts. Within projects focused on long-term simulations or projections, Regional Climate Models (RCMs) currently operate at horizontal grid resolutions between 25 and 50 km [e.g., PRUDENCE (Christensen and Christensen, 2007), ENSEMBLES (Hewitt, 2005) and NARCCAP (http://www.narccap.ucar.edu/)]. On the whole, there is an increasing demand of fine-scale data (e.g., Jeffrey et al., 2001; Huld et al., 2006; Hancock and Hutchinson, 2006; Daly et al., 2008; Tang et al., 2012) that can be useful to understand any environmental process linked to climate. Within the proper climatic context, a horizontal resolution of 7–10 km is currently recognized as a good target (e.g., Suklitsch et al., 2011).

Our research activity (e.g., Lanfredi et al., 2004; Simoniello et al., 2008, 2011) examines complex processes linking climate and the land surface (Piao et al., 2006; Cleland et al., 2007; Prieto-Blanco et al., 2009). Such studies use remote sensing observations of land and require realistic and accurate climatic surfaces obtained by interpolating data from meteorological stations to be interfaced with remote information. In particular, we need to construct air temperature surfaces that can be linked to land surface maps to

* Corresponding author. Tel.: +39 0971427284; fax: +39 0971427271.
E-mail address: lanfredi@imaa.cnr.it (M. Lanfredi).

http://dx.doi.org/10.1016/j.envsoft.2014.12.009
1364-8152/

better understand biosphere spatio-temporal patterns and to characterize exchange processes (e.g., carbon emission–absorption) that actively involve climatic fluctuations (e.g., Cox et al., 2000; Yuan et al., 2010).

Surface data are mostly obtained by the pure interpolation of sampled observations (e.g., by Thin Plate Smoothing Splines; Hopkinson et al., 2012). More complex strategies combine human-expert knowledge and statistical methods to satisfy the increasing demand for spatial climate data sets in digital form (Daly et al., 2008). All of these methodologies directly supply end-users with gridded data; the underlying physical mechanisms that shape the climatic surfaces are not singled out and thus remain encapsulated within the complexity of the gridding algorithms. Nevertheless, modelling the relationships between a given variable and the factors that generate its spatial patterns is crucial in many scientific frameworks. In our case, we have to consider that the spatial variability of both the land surface and low atmosphere variables is influenced by geography and topography. Any study focussing on fluctuations generated by mutual interactions between these two environments needs to discriminate geographic-induced background patterns that could distort correlation analyses.

This requirement led us to work on the development of a regressive approach that can account for causal linkages between geographic factors and temperature. General non-linear regression implies that functional form selection, estimation of best-fit parameters, and evaluation of fit performances are rather difficult. In contrast to linear regression, there is no closed-form expression for the best-fitting parameters and departures from the optimal approximation can occur, which could not be accounted for by global cost functions and require weighty goodness-of-fit tests (Caouder and Huet, 1997; Crainiceanu and Ruppero, 2004; Demidenko, 2006).

Here, we focus on a simpler approach by developing an additive regression model that is non-linear in the explanatory variables. The ability of such a model to generate random errors starting from spatially structured patterns can be considered as an *a posteriori* criterion to evaluate its performance. The main idea of our proposal is that we can use variogram analysis (Cressie, 1993; Wackernagel, 2003) to characterize the scale properties of the response variable along pseudo-directions that are defined by the explanatory variables of the model within an identification–estimation-checking iterative approach to model building. This analysis can be particularly useful in the diagnostic checking phase to verify the ability of the fit to remove correlation structures from the data and thereby randomize residuals from the fitted model ("whitening"). Efficient best fits should flatten the variogram at the right variance level; improper best fits should result instead in residual correlation between the response and explanatory variables over large scales. This validation is also important because it enables us to evaluate if the prediction error is actually the minimum allowed by the intrinsic degree of randomness of the data. Rigorously speaking, long-range correlation could also be observed in the case of fractal data, but this peculiar circumstance is recognizable due to the typical power law dependence that characterizes them (e.g., Brown and Liebovitch, 2010). Thus we are limited to consider determinism against stationary randomness. Of course, differently from the standard geostatistical applicative framework, the model variables are not necessarily spatial coordinates.

We illustrate our strategy by building up a geographical model for the climatic average of atmospheric temperature over Europe. Data from 387 meteorological stations were recorded over the 30-year period from 1961 to 1990, where the latest global "Normals" are currently defined for climate reference (http://www.wmo.int/pages/themes/climate/statistical_depictions_of_climate.php) according to the World Meteorological Organization. Although air surface temperature is one of the most continuous and studied variables within climate analyses, not only its deep dynamical features in time are still discussed (e.g., Lanfredi et al., 2009 and references therein) but also in truly applicative contexts there is no single strategic approach to the modelling, as observed above for climatic variables in general. We refer to the 8 km × 8 km resolution of the GIMMS-AVHRR (Global Inventory Modelling and Mapping Studies-Advanced Very High Resolution Radiometer) data, which are usually exploited for monitoring land cover in climatic studies (e.g., Zeng et al., 2013). This resolution corresponds well to the typical finest scales of RCMs (e.g., Suklitsch et al., 2011) and, as it will be shown in the following, emerges naturally from scale analyses as a reasonable boundary between locality and globality. The main variables shaping the basic structural part of the spatial variability of near-surface temperature at that resolution in a climatic context are latitude, longitude and elevation. We have also included the distance from the coastline to illustrate our approximation process step by step. The final part of the paper concerns a detailed discussion of the residuals from the fitted model and the comparison between the performances of our model against a standard multi-regressive linear model.

2. Data and study area

The annual mean air temperature data concerning the 30 years climatic period from 1961 to 1990 were obtained from 387 meteorological stations located in the European part of the Eurasian continent (Fig. 1) by averaging daily data. Most of the data were provided by the European Climate Assessment & Dataset (ECA&D) project (Klein-Tank et al., 2002); few stations (<2%) were integrated from local databases to introduce additional information in poorly represented areas.

Differences in latitude and elevation are expected to play a major role in determining the mean annual value of the air temperature, but the longitude and distance to the sea could also be significant parameters. In particular, from the point of view of general atmospheric circulation, the investigated area falls in the Ferrel cell of the Northern hemisphere where prevailing winds are westerlies. Because the west coast of Europe is located on the Atlantic Ocean, whereas the eastern part is continental, the westerlies move hot air masses inland from the sea in the direction of increasing longitude during winter. As a consequence, non-stationary behaviours are expected in the West(Sud)/East(Nord) direction. This variability should prevalently concern annual excursions, but the annual mean values could also be affected. Moreover, sea proximity, in general, modifies the minimum temperature in coastal swaths, which is why this parameter is included in the set of geographical parameters potentially involved in determining air temperature spatial variability.

3. Method

3.1. Variogram analysis

In this Section, we provide some basic definitions and concepts concerning the variogram analysis (for a detailed discussion see Cressie, 1993).

If $Z(\mathbf{s})$ is a regionalized stationary variable with a constant mean μ and variance σ^2 in a d-dimensional Euclidean space D, the quantity $2\gamma(.)$, which has been called a *variogram* by Matheron (1962), is defined as:

$$2\gamma(\mathbf{s_1}-\mathbf{s_2}) \equiv \mathrm{var}(Z(\mathbf{s_1})-Z(\mathbf{s_2})) \quad \text{for all } \mathbf{s_1},\mathbf{s_2} \in D \tag{1}$$

Due to the stationary assumption, this is a function of the increments $\Delta\mathbf{s} = \mathbf{s_1} - \mathbf{s_2}$ only and $\gamma(\Delta\mathbf{s}) \propto \sigma^2$ for large values of $\Delta\mathbf{s}$ asymptotically.

When the mean is assumed to be a constant, this equality holds:

$$\mathrm{var}(Z(\mathbf{s}+\Delta\mathbf{s})-Z(\mathbf{s})) = E(Z(\mathbf{s}+\Delta\mathbf{s})-Z(\mathbf{s}))^2 \quad \forall \mathbf{s}, \Delta\mathbf{s} \tag{2}$$

where $E(.)$ indicates the expected value, and we can estimate the variogram as:

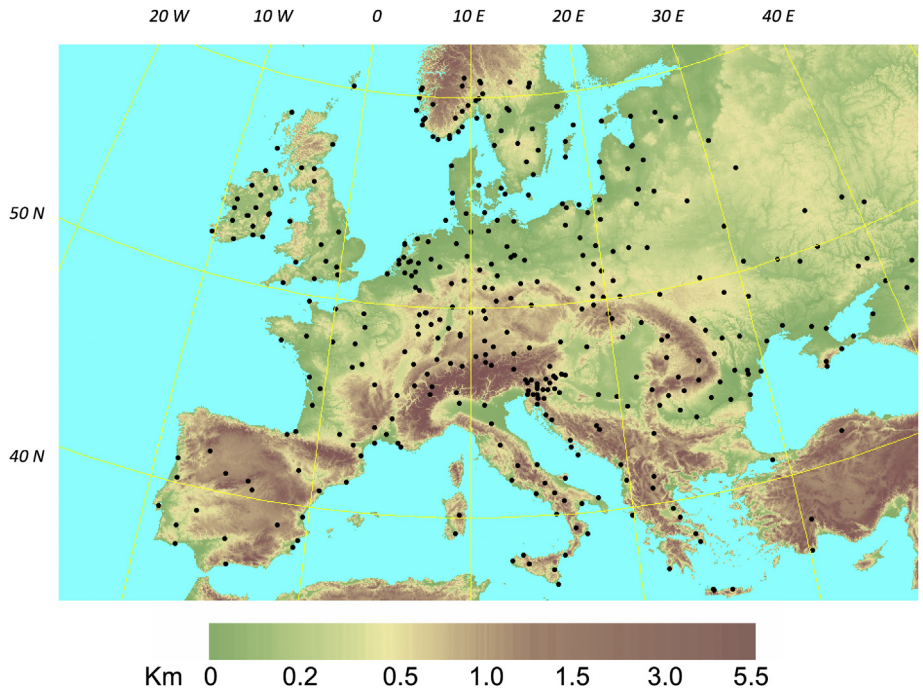

Fig. 1. Temperature database. Filled circles indicate the meteorological stations where observational data were recorded. Latitude ranges from 35° to 61° N, whereas longitude ranges from −10° to 42° E. The stations are superimposed on the Digital Elevation Model of the region derived from 90 m SRTM (Shuttle Radar Topography Mission) data available at http://srtm.csi.cgiar.org/.

$$2\hat{\gamma}(\Delta\mathbf{s}) = E(Z(\mathbf{s}+\Delta\mathbf{s}) - Z(\mathbf{s}))^2 \quad (\mu = const.) \tag{3}$$

Eq. (3) is often reported as the variogram definition in applicative contexts. Hereafter, we will use the word *variogram* to indicate the statistics defined in Eq. (3).

Let us consider now a non-stationary variable $Y(\mathbf{s})$ that can be represented as the superposition of a surface trend $\mu(\mathbf{s})$ and a stochastic variable $\sigma Z(\mathbf{s})$, where $Z(\mathbf{s})$ is a normal variable, as is usually the case in many applicative fields, particularly in climate problems:

$$Y(\mathbf{s}) = \mu(\mathbf{s}) + \sigma Z(\mathbf{s}), \quad \mathbf{s}\in D \tag{4}$$

In this case, the second term of Eq. (3) can be written as:

$$E(Y(\mathbf{s}) - Y(\mathbf{s}+\Delta\mathbf{s}))^2 = 2\gamma_Z(\Delta\mathbf{s}) + (\mu(\mathbf{s}) - \mu(\mathbf{s}+\Delta\mathbf{s}))^2 \tag{5}$$

and thus, on scales larger than that characterizing stochastic correlation, we have:

$$\hat{\gamma}(\Delta\mathbf{s}) \propto \sigma^2 + \frac{(\mu(\mathbf{s}) - \mu(\mathbf{s}+\Delta\mathbf{s}))^2}{2} \tag{6}$$

Possible systematic behaviours, resulting in large-scale correlation, appear within such empirical variograms as trends steadily climbing beyond the total variance of the data. In standard applications, these trends are preventively removed to study the small-scale properties of the process, but it is possible to use variogram analysis to examine such surface trends.

3.2. Rationale

Because trend and local variability are generally on rather different scales, we can separate the stochastic and deterministic terms in Eq. (6) by inspecting the sampling variogram. The deterministic trend does not contribute significantly to the variogram at the local scale due to its continuity properties ($\lim_{\Delta\mathbf{s}\to 0} (\mu(\mathbf{s}) - \mu(\mathbf{s}+\Delta\mathbf{s}))^2 = 0$), and by looking at such small scales, we can obtain an estimate of the variance σ^2 not explained by the trend. Thus the actual reference variance σ_F^2 for evaluating the adequacy of the fit is:

$$\sigma_F^2 = \hat{\gamma}(\Delta\mathbf{s}) - \sigma^2 \tag{7}$$

This variance is obviously lower than that of the data. As an example, the maximum determination coefficient we can obtain is not $R_{\max} = 1$ but lower:

$$R_{\max}^2 = \sigma_F^2 \big/ \sigma_Y^2 \tag{8}$$

Thus, Eq. (8) gives us a criterion to evaluate the explained variance. In addition, variogram analysis of residuals permits a quick and easy detection of the improper functional form, capturing subtle deviations from the flat behaviour expected for well-fitted surface trends. Generally, geographical coordinates are used in classical applications of geostatistics to extract spatial correlation structures. Here, we instead use the selected explanatory variables of the fitting model to infer the main systematic patterns in the long range. In practice, we define the vector of the explanatory variables $\mathbf{x} = (x_1, ..., x_n)$ and search for an additive model $Y(\mathbf{x}) = X_1(x_1) + ... + X_n(x_n) + \varepsilon$, where $Y(\mathbf{x})$ is the modelled variable; the terms $X_i(x_i),(i = 1 \ldots n)$ are analytical, possibly non-linear functions; and ε are stochastic residuals. We estimate variograms in the n pseudo-directions defined by each one of the variables $(x_1, ..., x_n)$ within a step-wise procedure to evaluate the expected fit error and the actual ability of selected functional forms to reconstruct the actual long-range patterns. A final standard variogram analysis is performed on ε to assess the randomness of the residuals on the scales concerned thus making the method auto-consistent.

According to the scale separation approach, this analysis allows us to model broad scale patterns. For continuous variables, a local scale correlation is expected to be present in the residuals. If the correlation range is comparable to the typical distances between the meteorological stations, the variability generated at these resolutions can be successively interpolated by using geostatistical tools (e.g., by Kriging, Cressie, 1993).

4. Results

4.1. Development of the regressive model

The geographical temperature distribution is likely the main source of the large-scale coherence of our data. To express this regularity as an explicit function of geographical variables, we looked at the plots of T versus the explanatory variables (latitude, elevation, longitude, and distance from the coastline).

In Fig. 2, the major roles of latitude (Fig. 2a) and elevation (Fig. 2b) are evident. Nevertheless, the non-uniform distribution of mountain chains implies an accumulation of low temperature values at latitudes between 37° and 50° N (Fig. 2a) that can distort the estimation of latitude dependence. Thus, the first problem was to convert the temperature to the temperature at sea level. The dependence of temperature on elevation has been widely discussed in the literature, which reports a linear average temperature decrease with a lapse rate around $\Delta = 6$ °C km^{-1}, with slight

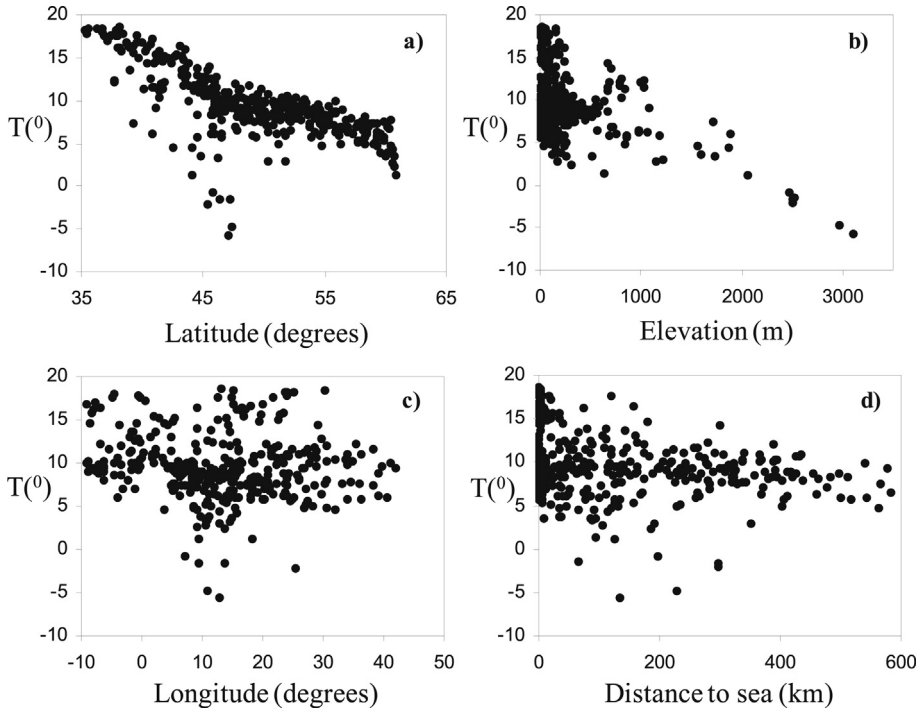

Fig. 2. Plots of the temperature data against: a) latitude; b) elevation; c) longitude; d) distance from the coastline.

seasonal differences (e.g., Linacre, 1992; Hudson and Wackernagel, 1994; Minder et al., 2010). If we would estimate such a lapse rate directly from the plot in Fig. 2b we would obtain an incorrect value due to the spread of temperature values in low elevation sites, which are mainly controlled by latitude. To minimize the influence of latitude, we sorted the sampling sites according to increasing latitude and focused on elevation (Δh) and temperature (ΔT) gradients. In such a way, we looked at elevation increments between sites located on the same or close latitudinal parallels. The variogram of ΔT against Δh (Fig. 3a) shows that the expected percentage of unexplained variance is approximately 12–14 % which implies $R^2_{max} = 0.88 - 0.86$.

Such a variance is prevalently due to other factors because temperature variations with elevation become significant at approximately $\Delta h = 100$ m, where the trend becomes evident (Fig. 3a). Over a larger scale, linearity seems to work rather well (Fig. 3b). The estimated lapse rate $\alpha = 5.6\ ^\circ C\ km^{-1}$ is similar to the values reported in the literature and the squared determination coefficient is $R^2 = 0.87$, which is consistent with R^2_{max}.

Once temperature is converted to the sea level temperature (T_S), we can model it according to latitude (Fig. 4a). The variogram in Fig. 4b reveals that approximately 7% of the variance can be considered independent of latitude, resulting in the reference determination coefficient of $R^2_{max} = 0.93$. In addition, it shows that the sea level mean temperature can be considered stationary over a latitude range $\Delta lat = 1^\circ$. This scale signifies the first crossover (from independence to dependence) that characterizes the relationship between temperature and latitude.

To model the data in Fig. 4a, we consider three fitting functions that will be discussed comparatively: the linear trend, the second order polynomial trend, and the Gaussian trend.

4.1.1. Linear trend

Linear approximation is usually adopted when the target area is rather limited in latitude (e.g., Chuanyan et al., 2005; Shao et al., 2012).

$$T_L(lat) = a_L lat + b_L \tag{9}$$

Whatever the actual decay of temperature with latitude may be, T_L should be suited when the Earth surface portion that we consider can be approximated with a tangential plan. To the best of our knowledge, generally explicit cross-over scales between linearity and non-linearity are not discussed.

4.1.2. Second order polynomial trend

Linacre and Geerts (Linacre, 1992; Linacre and Geerts, 2002) found that a second order power law well represented the latitude trend in data concerning the whole Northern Hemisphere.

According to the authors, this model is grounded on physical reasons because it is the consequence of the influence of the solar irradiance at the ground. The solar irradiance at the ground depends on the product of the irradiance at the top of the atmosphere and the attenuation of the solar beam through the atmosphere, both depending on latitude (Linacre, 1992). We tested a general second order polynomial (Adjusted Linacre and Geerts trend) assuming that the quadratic dependence due to solar irradiance has to be taken into account. However, the approximated model has to include some additional terms to account for some peculiar features of the European continent, in particular, the land–sea distribution (this influences temperature and is also expected to introduce a slight dependence between latitude and longitude that are mutually constrained within the ideal equation that would define the European land surface):

$$T_{ALG}(lat) = a_{ALG} lat^2 + b_{ALG} lat + c_{ALG} \tag{10}$$

4.1.3. Gaussian trend

Upon careful inspection of Fig. 4a (see red lines), three apparent regimes roughly coincide with the Mediterranean area, the continental area, and the northernmost maritime swath. The

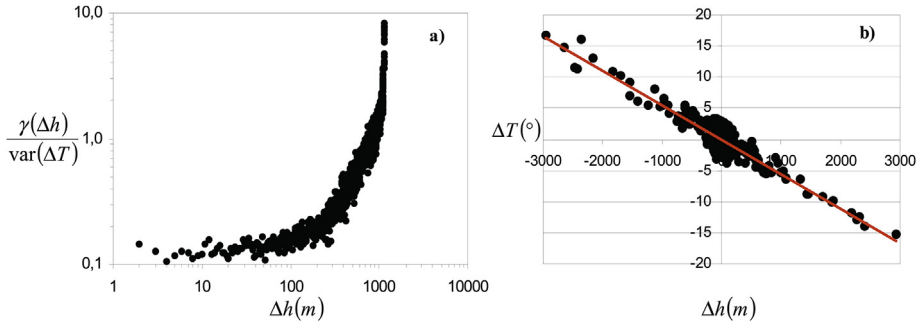

Fig. 3. Temperature gradients vs. elevation gradients for the estimation of the lapse rate α: a) variogram; b) data plot. Linearity is rather clear even if the variability of low elevation temperature is strongly affected by other explanatory variables that determine a cloud of points around the origin of the plot. Red line shows the best linear fit $\Delta T = -0.0056$, $\Delta h = -0.0354$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

superposition of two Gaussian functions could account for slight curvature changes and the Gaussian parameters could inform us about temperature variability features within these areas:

$$T_G(lat) = A_1 e^{-\frac{(lat-b1)^2}{c_1^2}} + A_2 e^{-\frac{(lat-b2)^2}{c_2^2}} \quad (11)$$

Fig. 5 shows variograms for the three selected models. The value of R^2 associated to the linear fit is $R^2 = 0.89$ with approximately 11% of unexplained variance. The linear fit regularizes data well in a range of approximately 2 or 3 latitude degrees, but is inefficient for larger areas. The estimated unexplained variance is approximately 8.4% for the ALG filter, which regularizes better temperature data. The Gaussian filter with two components shows the best performance because the estimated unexplained variance (~7.4%) is close to the expected one.

The Gaussian solution follows the curvature changes that we observe along the temperature pattern (see Fig. 6). It is interesting to note that the two Gaussian distributions are centred in the southern part of the two maritime swaths, whereas the continental pattern emerges from the superposition of these distributions that are weighted by the coefficients $A_1 \cong 2A_2$. Surely these patterns are grounded on general physical reasons, but nevertheless, we hypothesize that a small contribution could result from the different distributions of longitude of the three areas. The middle zone is the most continental, as it includes the main part of the inland eastern area, whereas the westernmost part of the grid in Fig. 1 is occupied by sea. As a consequence, we presume that the non-linear fit partially accounts for latitude–longitude constrains that are imposed by the specific land mass geography of the European peninsula.

We examined air temperature dependence on longitude and distance from the coastline by following the same approach. A slightly decreasing drift with longitude was detected in the residuals obtained by filtering dependence on latitude. Although the descriptive power of the linear best fit ($R^2 = 0.60$) was rather poor due to the high noise content, we decided to include this drift in the model because R^2_{max} was estimated to be approximately 0.5–0.7. No significant correlation with the distance from the coastline was found.

The final model was:

$$T(h, lat, lon) = -0.0056h - 0.05lon + 16.94e^{-\frac{(lat-35.37)^2}{(10.3)^2}} + 8.06e^{-\frac{(lat-52.72)^2}{(12.52)^2}} + 0.68 \quad (12)$$

The parameterization of Eq. (12) was performed by integrating the contribution of all of the meteorological stations. Nevertheless, the actual predictive skill of the model was obtained by computing mean temperature in each of the 387 meteorological stations with parameters estimated on the basis of the remaining 386 stations (Cross-Validation). The unexplained variance was approximately 3% with an RMSE = 0.7 °C.

The final variograms (Fig. 7), estimated on the residuals of the model in Eq. (12), demonstrate no significant patterns, and Fig. 8 shows the synthetic map of the mean climatic air temperature. It is possible to identify latitude patterns and the significant influence of elevation that lowers temperature. Slight West–East behaviours are also detectable.

4.2. Analysis of the residuals. Comparison with a linear regression model

According to our results, the final representation of the temperature T in a site $\boldsymbol{x}$ is:

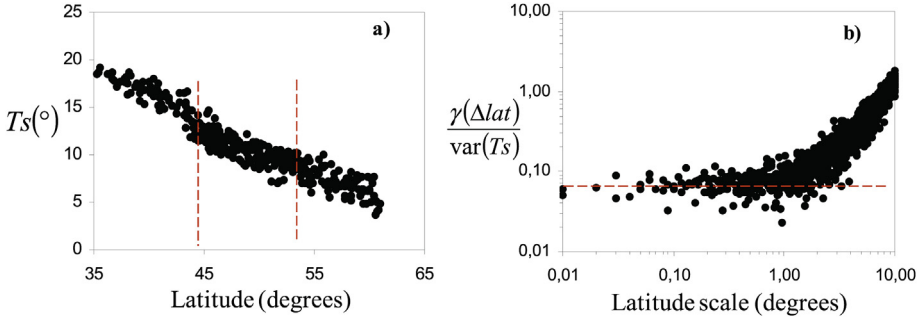

Fig. 4. Sea level temperature (a) and related variogram against latitude difference (b); the red lines in (a) mark slight discontinuities, and the red line in (b) marks the expected unexplained variance. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

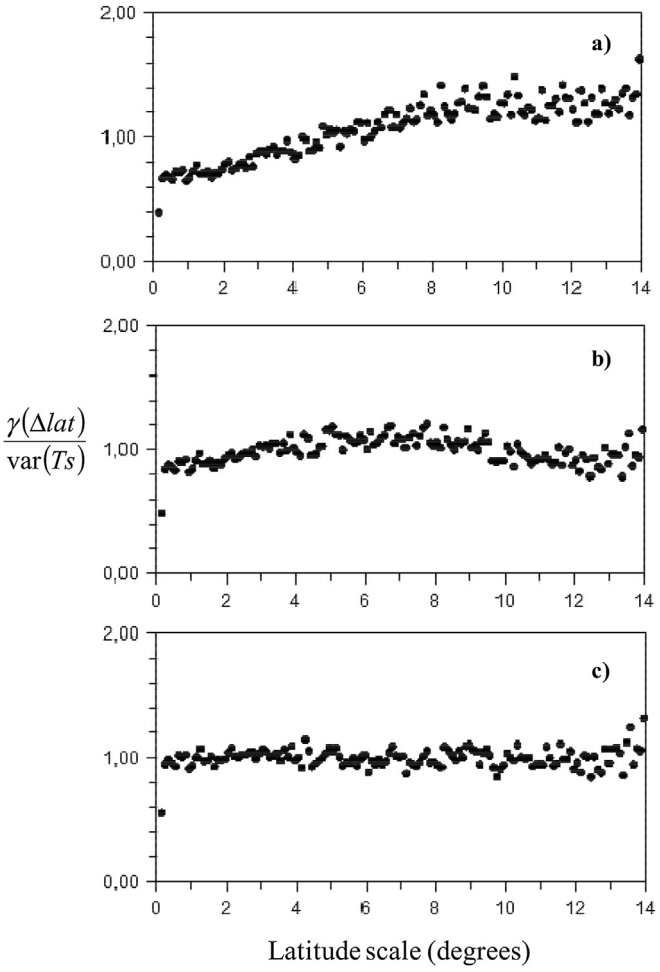

Fig. 5. Variogram of the residuals from three fits: a) linear ($T_L(lat) = -0.5lat + 36.2$); b) second order polynomial ($T_{ALG}(lat) = -0.01lat^2 - 1.6lat + 63.2$); c) Gaussian superposition $T_G(lat) = 16.94e^{-\frac{(lat-35.37)^2}{(10.3)^2}} + 8.06e^{-\frac{(lat-52.72)^2}{(12.52)^2}}$.

$$T(\boldsymbol{x}) = T(lat, lon, h) + \varepsilon \tag{13}$$

where $T(lat, lon, h)$ is obtained from the model in Eq. (12) and the variable ε represents the residual component, as explained in Section 3.2.

To estimate the local spatial correlation not explained by our model, we analysed these residuals in detail.

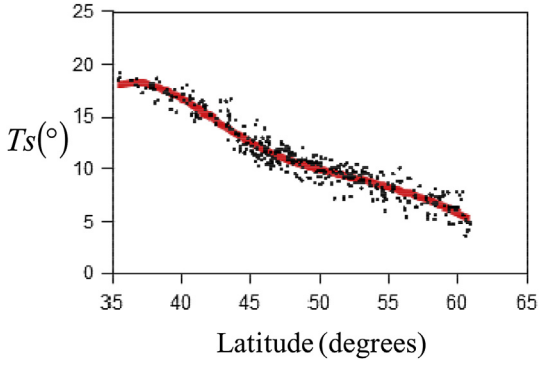

Fig. 6. Sea level temperature vs. latitude. The red line shows the Gaussian filter selected for approximating data. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

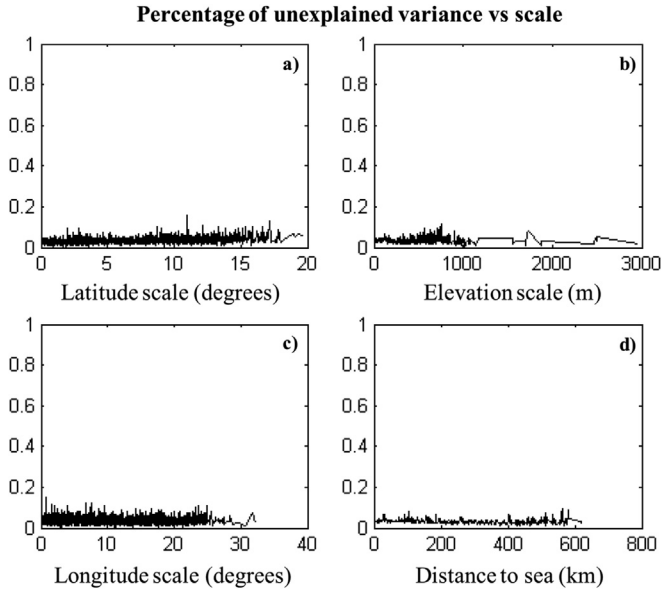

Fig. 7. Variograms of the residuals from model (Eq. (12)) vs. the explanatory variables. As expected, although distance to sea was excluded by the model, detrending also lowers variance in d).

The box-plot in Fig. 9 summarizes the distribution of ε, which is symmetric and well centred around the value $\varepsilon = 0$, with most of the values located in the range $(-0.5°, 0.5°)$. A dozen values are recognized as outliers. Incidentally, our procedure is not particularly sensitive to outliers because it is not a proper interpolation tool that tries to "pass" through all of the sample points. The largest residuals are obtained in the Mediterranean region, at the limit between Mediterranean and continental climates, and are likely related to the complex patterns of cold polar winds that are driven by mountainous barriers in a prevalently mild zone. Large positive residuals are also obtained in hot urban areas (e.g., Paris, Prague). These outliers are mainly responsible for the estimated RMSE = 0.7 °C. If we eliminate the Mediterranean area (up to 46–47 °N), the value of this error drops down to RMSE = 0.5 °C. These errors are comparable with errors reported in the literature and obtained in the analysis of temperature "Normals" through complex interpolation tools (see Hopkinson et al., 2012).

To bring out the added value of our procedure, we fitted a linear multi-regressive model (with the same variables *lat*, *lon*, and *h*) to the data. Fig. 10 shows the variogram of the residuals obtained by applying our model (black line) and the linear model (red line). The two variograms roughly coincide up to about a few hundreds of kilometres. Although the estimated RMSE = 0.85 °C of the linear model is not particularly worse than that of our model, it does not represent a measure of the adequacy of the fit because the linear model is not able to make the data stationary on large scales and provides ever more inefficient predictions. In this case, the mere deterministic approximation does not explain all of the variance generated by the broad scale coherence. To account for all of the variance, such residuals should be further interpolated, as an example through Kriging, thus splitting into two components the variability accounted for by the only model in Eq. (12).

On the contrary, the variogram of the residuals of our model reaches the expected percentage of total variance (0.03) in a range of about ten kilometres (our target resolution). Our model is an accurate descriptor of the climatic field in grid cells of 10 km × 10 km and the short correlation range of the local fluctuations are a direct consequence of the good trend removal. This

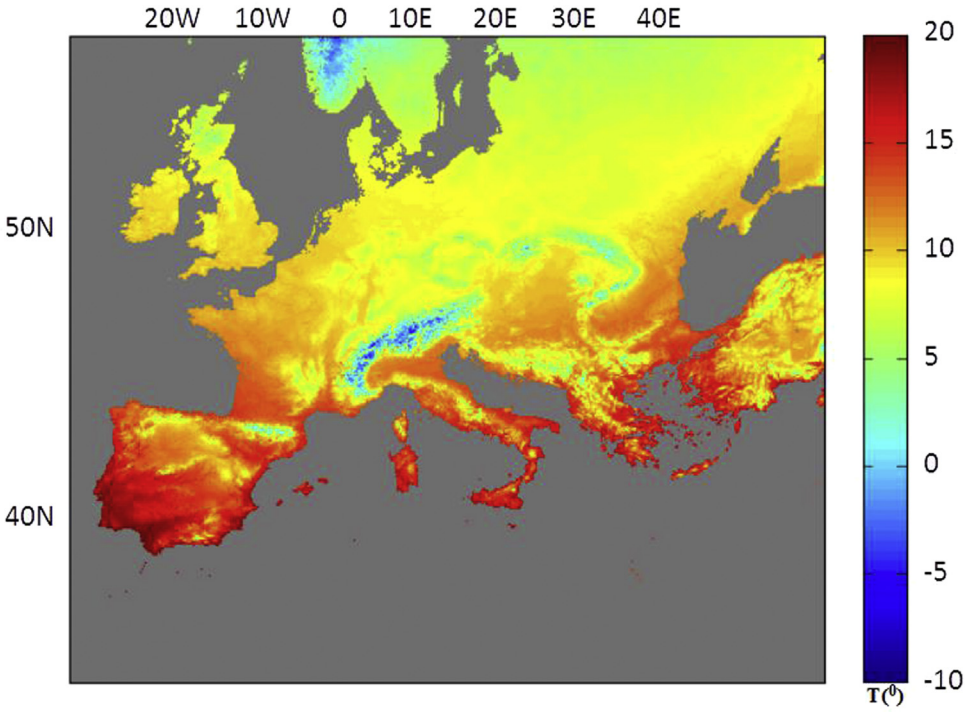

Fig. 8. Synthetic map of the mean climatic air temperature estimated through model in Eq. (12) using an 8 km × 8 km grid.

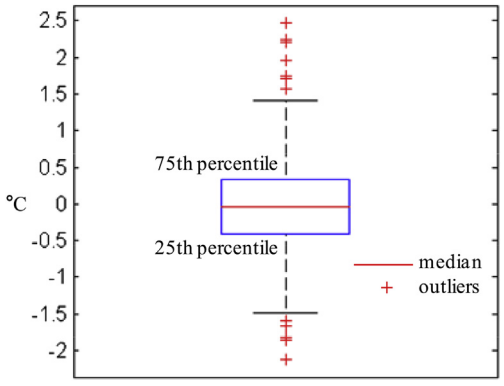

Fig. 9. Box-plot of the model residuals in Eq. (12).

scale, which naturally emerges from our analysis, as a reasonable boundary between locality and globality, is consistent with the typical finest scale RCMs (e.g., Suklitsch et al., 2011). The short correlation range that marks the representative domain of a single station (on grid cells wider than 10 km resolution are non-correlated) implies that a dense sampling dataset would be necessary to increase resolution. A rough estimation for the European continent (~10,000,000 km^2) recommends approximately 100,000 stations.

We also evaluated the possibility of introducing parameters accounting for very local topographic details, which have provided important predictive improvements in mountainous areas (Minder et al., 2010). Nevertheless, the parameters we can obtain from a DEM, such as the slope and aspect, are not able to improve the model due to the presence of heterogeneity, poor high-elevation sampling over Europe and the lack of detailed local-scale information.

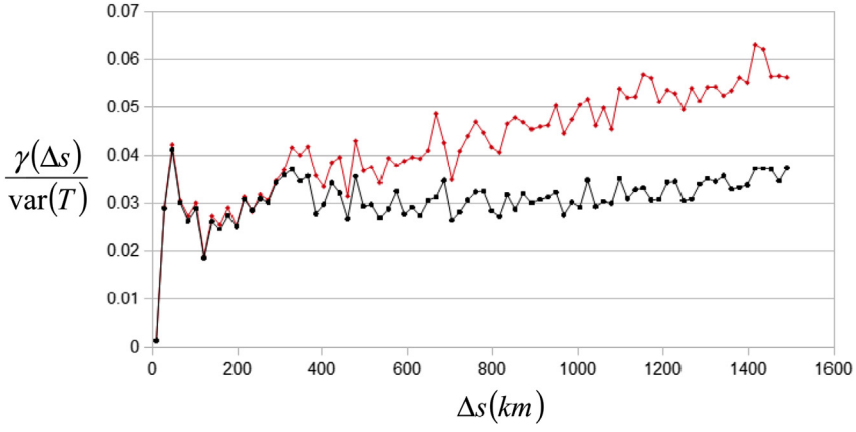

Fig. 10. Sample variograms of the residuals of two deterministic models: the model in Eq. (12) (black line); the linear regression model $T = -0.0057h - 0.53lat - 0.063lon + 37.73$ (red line). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

5. Conclusions

We have proposed a regressive method for modelling broad-scale climate fields in a general non-linear context. Model building is achieved by a three-stage procedure based on identification, estimation, and diagnostic checking. In both the first and last stages, we use variogram analysis in the domains defined by each explanatory variable. In the exploration phase, the variogram informs us about the correlation between a response variable and a candidate explanatory variable. In particular, the variogram allows us to infer the scales involved and the level of the variance that cannot be explained. In the diagnostic phase, we analyse the same scale properties after the removal of the candidate fit. A model is acceptable if the residuals from the fitted model are consistent with a random field on the scales concerned and the variance is reduced to the previously estimated level. Because scales are separated, this procedure is also useful for identifying scale ranges where simple approximations may work well.

The application of the method to the climatic mean temperature recorded over Europe shows good results, especially considering the low number of explanatory variables (elevation, latitude, and longitude). A particularly interesting result was found about the latitude-dependence of temperature. While it is common knowledge that solar radiation is unevenly distributed and that intensity varies from one location to another depending upon the latitude, it is more difficult to understand how these patterns are altered by the land mass distribution. The best fit we designed describes three different climatic sub-zones in the European area: the Mediterranean zone, the continental area, and the northern maritime swath. Such a partition, which accounts for two relative maxima in the maritime areas, reflects the synergic effect of radiance variability against latitude and land–sea distribution, which is characterized by the presence of a massive area of land in the continental zone. Over land areas that do not exceed 2–3 latitude degrees, a linear approximation is adequate.

A range of approximately 10 km marks the decay of the local correlation of the residuals that cannot be further interpolated due to the lack of information at such fine scales. The unexplained variance of our model is approximately 3% of the total variance; an estimate of the error by cross validation is RMSE = 0.7 °C (RMSE = 0.5 °C for the continental area). Such a prediction error is only due to the peculiar microclimatic variability, especially in the Mediterranean region and comes from an intrinsic descriptive limit of the data. The final diagnostic check by traditional variogram analysis enables us to evaluate the overall representation adequacy of our model.

The result of our procedure is not only simple, parsimonious, and able to synthesize the effects of the mechanisms that shape the climatological temperature near the terrestrial surface, but is also self-consistent because the ability to randomize the data gives an intrinsic value to the model, independent of the building process.

Acknowledgements

This work is inserted in the framework of the project ORIENTGATE – A network for the integration of climate knowledge into policy and planning-funded by the South East Europe Transnational Cooperation Programme (CCI 2007CB163PO069).

References

Brohan, P., Kennedy, J.J., Harris, I., Tett, S.F.B., Jones, P.D., 2006. Uncertainty estimates in regional and global observed temperature changes: a new dataset from 1850. J. Geophys. Res. 111, D12106.

Brown, C., Liebovitch, L., 2010. Fractal Analysis. SAGE Publication Inc., Los Angeles, CA, 90 pp.

Caouder, N., Huet, S., 1997. Testing goodness-of-fit for nonlinear regression models with heterogeneous variances. Comput. Stat. Data Anal. 23, 491–507.

Christensen, J.H., Christensen, O.B., 2007. A summary of the PRUDENCE model projections of changes in European climate by the end of this century. Clim. Change 81, 7–30. http://dx.doi.org/10.1007/s10584-006-9210-7.

Chuanyan, Z., Zhongren, N., Guodonga, C., 2005. Methods for modelling of temporal and spatial distribution of air temperature at landscape scale in the southern Qilian mountains, China. Ecol. Model. 189, 209–220.

Cleland, E.E., Chuine, I., Menzel, A., Mooney, H.A., Schwartz, M.D., 2007. Shifting plant phenology in response to global change. Trends Ecol. Evol. 22, 357–365.

Cox, P.M., Betts, R.A., Jones, C.D., Spall, S.A., Totterdell, I.J., 2000. Acceleration of global warming due to carbon-cycle feedbacks in a coupled climate model. Nature 408 (6809), 184.

Crainiceanu, C.M., Ruppero, D., 2004. Likelihood ratio tests for goodness-of-fit of a nonlinear regression model. J. Multivar. Anal. 91, 35–52.

Cressie, N.A.C., 1993. Statistics for Spatial Data. John Wiley, New York, 928 pp.

Daly, C., Halbleib, M., Smith, J.I., Gibson, W.P., Doggett, M.K., Taylor, G.H., Curtis, J., Pasteris, P.P., 2008. Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. Int. J. Climatol. 28, 2031–2064. http://dx.doi.org/10.1002/joc.1688.

Demidenko, E., 2006. Criteria for global minimum of sum of squares in nonlinear regression. Comput. Stat. Data Anal. 51, 1739–1753.

Feser, F., Burkhardt, R., von Storch, H., Winterfeldt, J., Zahn, M., 2011. Regional climate models add value to global model data: a review and selected examples. Bull. Am. Meteorol. Soc. 92, 1181–1192. http://dx.doi.org/10.1175/2011BAMS3061.1.

Geiger, R., Aron, R.H., Todhunter, P., 2003. The Climate Near the Ground. Rowman and Littlefield Publishers, Lanham, MD, USA, 584 pp.

Hancock, P.A., Hutchinson, M.F., 2006. Spatial interpolation of large climate data sets using bivariate thin plate smoothing splines. Environ. Model. Softw. 21, 1684–1694.

Hewitt, C.D., 2005. The ENSEMBLES project: providing ensemble-based predictions of climate changes and their impacts. EGGS Newsl. 13, 22–25.

Hopkinson, R.F., Hutchinson, M.F., McKenney, D.W., Milewska, E.J., Papadopol, P., 2012. Optimizing input data for gridding climate normals for Canada. J. Appl. Meteor. Climatol. 51, 1508–1518.

Hudson, G., Wackernagel, H., 1994. Mapping temperature using kriging with external drift: theory and an example from Scotland. Int. J. Climatol. 14 (1), 77–91. http://dx.doi.org/10.1002/joc.3370140107.

Huld, T.A., Suri, M., Dunlop, E.D., Micale, F., 2006. Estimating average daytime and daily temperature profiles within Europe. Environ. Model. Softw. 21, 1650–1661.

Jeffrey, S.J., Carter, J.O., Moodie, K.B., Beswick, A.R., 2001. Using spatial interpolation to construct a comprehensive archive of Australian climate data. Environ. Model. Softw. 16, 309–330.

Jones, G.S., Stott, P.A., 2011. Sensitivity of the attribution of near surface temperature warming to the choice of observational dataset. Geophys. Res. Lett. 38, L21702.

Jones, P.D., Lister, D.H., Osborn, T.J., Harpham, C., Salmon, M., Morice, C.P., 2012. Hemispheric and large-scale land-surface air temperature variations: an extensive revision and an update to 2010. J. Geophys. Res. 117, D05127.

Klein-Tank, et al., 2002. Daily dataset of 20th-century surface air temperature and precipitation series for the European climate assessment. Int. J. Climatol. 22, 1441–1453.

Lanfredi, M., Simoniello, T., Macchiato, M., 2004. Temporal persistence in vegetation cover changes observed from satellite: development of an estimation procedure in the test site of the Mediterranean Italy. Remote Sens. Environ. 93 (4), 565–576.

Lanfredi, M., Simoniello, T., Cuomo, V., Macchiato, M., 2009. Discriminating low frequency components from long range persistent fluctuations in daily atmospheric temperature variability. Atmos. Chem. Phys. 9, 4537–4544.

Linacre, E., 1992. Climate Data and Resources: a Reference and Guide. Routledge, London, 366 pp.

Linacre, E.T., Geerts, B., 2002. Estimating the annual mean screen temperature empirically. Theor. Appl. Climatol. 71, 43–61.

Matheron, G., 1962. Traité de Géostatistique Appliquée. In: Mémoires du Bureau de Researches Géologiques et Minières, n.14, Tome I. Éditions Technip, Paris.

Minder, J.R., Mote, P.W., Lundquist, J.D., 2010. Surface temperature lapse rates over complex terrain: lessons from the Cascade Mountains. J. Geophys. Res. 115, D14122. http://dx.doi.org/10.1029/2009JD013493.

Piao, S., Fang, J., Zhou, L., Ciais, P., Zhou, B., 2006. Variations in satellite-derived phenology in China's temperate vegetation. Glob. Change Biol. 12, 672–685.

Prieto-Blanco, A., North, P.R.J., Barnsley, M.J., Fox, N., 2009. Satellite-driven modelling of Net Primary Productivity (NPP): theoretical analysis. Remote Sens. Environ. 113 (1), 137–147.

Rummukainen, M., 2010. State-of-the-art with regional climate models. Wiley Interdiscip. Rev. Clim. Change 1 (1), 82–96.

Shao, J., Li, Y., Ni, J., 2012. The characteristics of temperature variability with terrain, latitude and longitude in Sichuan-Chongqing Region. J. Geogr. Sci. 22, 223–244.

Simoniello, T., Lanfredi, M., Liberti, M., Coppola, R., Macchiato, M., 2008. Estimation of vegetation cover resilience from satellite time series. Hydrol. Earth Syst. Sci. 12, 1053–1064.

Simoniello, T., Lanfredi, M., Coppola, R., Imbrenda, V., Macchiato, M., 2011. Correlation of vegetation and air temperature seasonal profiles: spatial arrangement and temporal variability. In: Zhang, X. (Ed.), Phenology and Climate Change. Intech Open Access Publisher, Rijeka, pp. 273–296.

Suklitsch, M., Gobiet, A., Truhetz, H., Awan, N.K., Goettel, H., Jacob, D., 2011. Error characteristics of high resolution regional climate models over the Alpine area. Clim. Dyn. 37 (1–2), 377–390. http://dx.doi.org/10.1007/s00382-010-0848-5.

Tang, L., Su, X., Shao, G., Zhang, H., Zhao, J., 2012. A Clustering-Assisted Regression (CAR) approach for developing spatial climate data sets in China. Environ. Model. Softw. 38, 122–128.

Wackernagel, H., 2003. Multivariate Geostatistics: an Introduction with Applications, third ed. Springer-Verlag, Berlin. 402 pp.

Yuan, W., Liu, S., Yu, G., Bonnefond, J.M., Chen, J., Davis, K., Desai, A.R., Goldstein, A.H., Gianelle, G., Rossi, F., Suyker, A.E., Verma, S.B., 2010. Global estimates of evapotranspiration and gross primary production based on MODIS and global meteorology data. Remote Sens. Environ. 114 (7), 1416–1431.

Zeng, F., Collatz, G.J., Pinzon, J.E., Ivanoff, A., 2013. Evaluating and quantifying the climate-driven interannual variability in Global Inventory Modeling and Mapping Studies (GIMMS) Normalized Difference Vegetation Index (NDVI3g) at global scales. Remote Sens. 5, 3918–3950. http://dx.doi.org/10.3390/rs5083918.