

Generating carbon flux model parameters and drivers via space/time geostatistics

Rationale:

As we know, our best moderate scale (1-10 Km²) estimates of the annual carbon budget of a vegetative surface come from flux tower measurements. However, these estimates are limited by the requirement of homogenous surface cover, and homogenous topography. The second requirement is due to slow air currents at night, which cause problems with energy budget closure by advection (at a roughness velocity below the sensitivity of the instrument). Thus measurement of carbon fluxes in mountainous terrain can be problematic. This is unfortunate, as mountain ecosystems display interesting and varied ecology; mostly because of agricultural limitations associated with the topography (i.e. 'natural' forests tend to exist on complex terrain).

One solution to the problem of estimating carbon budgets over such terrain is via ecosystem modelling. The problem here is that in order to arrive at sensible estimates, models must be adequately constrained with data. Whilst we are able to achieve adequate parameterisation at single points in space (e.g. around a flux tower), we are still faced with problems in areas which have perhaps more interesting ecological features, where data may be sparse. Over very large scales (global), 'spin-up' type simulation approaches allow quantification of fluxes, but over regional scales, assumptions of semi-equilibrium states most likely misrepresent the actual flux magnitude; to imply equilibrium at this scale is to ignore the ecology, management, etc.

Remotely sensed data offers a unique, large area data source to constrain models, and this remains an area of development (e.g. our data assimilation work). An alternative solution is to parameterise and drive models with ground based data from sparse measurement networks. One characteristic of such data is that it tends to be spatially sparse, yet temporally dense: a weather station may log hourly data, yet the average separation distance of the network may be tens of Km). The challenge here is to accurately estimate data values at un-sampled locations in space and time, and geostatistics provides a set of tools to achieve this.

However we choose to parameterise and drive our models, the problem remains that estimated surfaces are often poorly constrained, and have little or no explicit quantification of uncertainty: As such a modelling exercise can provide an estimate of the carbon budget over a large area, which is conditionally dependant on parameterisations of unknown accuracy, which represent only a single realisation of the stochastic process they intend to reproduce: Knowing the RMS error of a single surface does not allow us to identify areas where our final estimation is poor.

Geostatistical tools such as inverse distance weighting, thin plate splines, or, at a basic level Veronoi polygons allow us to make spatial averages, but do not quantify error in a spatially explicit manner. As such, I doubt whether traditional attempts to parameterise carbon models over moderate areas are even able to accurately determine the sign of the flux. For this reason, modern geostatistical methods are preferable fro parameterising and driving our models: Kriging allows us to more objectively interpolate our data over the domain of interest, and provides the basis for spatial error quantification via statistical simulation. This is of considerable benefit for quantifying regional carbon budgets, as an ensemble of equi-probable parameter sets can be derived from the marginal distribution (i.e. histograms of the data), and the known spatiotemporal autocorrelations. These states can then be propagated through the model to produce probabilistic estimates of the flux (i.e. we get a posterior PDF at each pixel). At the least, this allows us to put confidence intervals around our flux estimates, but can potentially identify areas in which our models perform poorly.

Current work:

My current work is centred on the production of meteorological driver surfaces; specifically at the moment, Tmin. We have a network of 113 met stations across Oregon, ranging from the East to the West of the central Cascade Mountains (Fig 1).

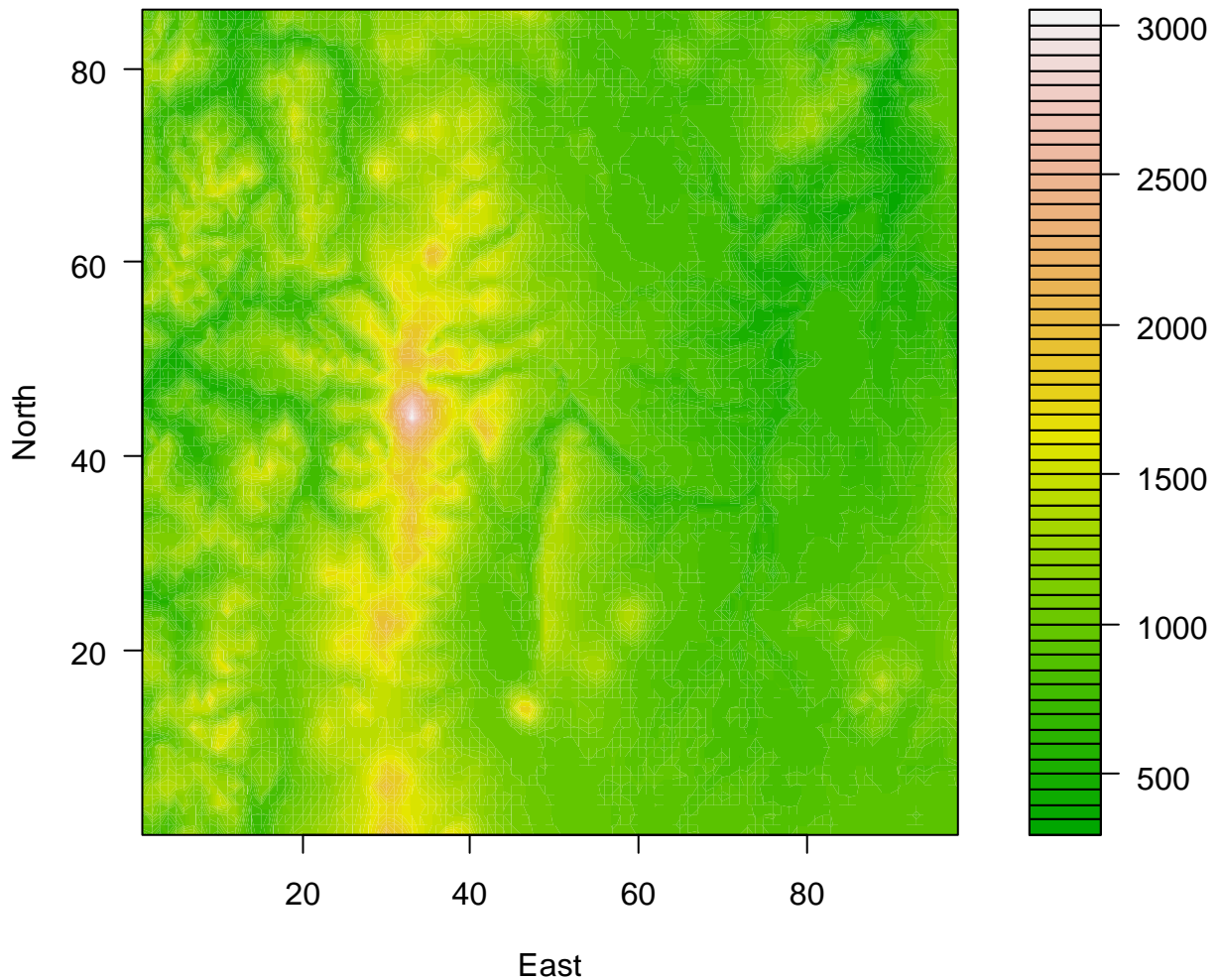


Fig 1: DEM of the study area; Eastern Cascades, Oregon.

I intend to produce the drivers by Kriging the data values to fill the domain, whilst imposing constraints on the spatial distribution of values through physical trends (e.g. temperature lapse rates with elevation). In order to adequately describe the method, I will have to briefly resort to theory (hopefully with a minimal tedium).

Kriging is essentially a regression, and can be expressed in the form:

$$y = M \cdot x + c$$

Where M is a matrix of weights and c is the global mean. Thus a prediction location has its value specified by a linear combination of the weighted contribution of the residuals of known data values and the global mean, in a way exactly analogous to an AR(x) model in time series analysis. Weights are derived from the statistical relationship of covariance against separation distance; things closer together are more similar, and are thus highly weighted. This separation of the spatial trend into a mean process, and a spatially correlated residual noise process is the rationale of Kriging. It is worth noting that in the above formulation, inverse distance weighting methods are a special case of the Kriging system; the difference is in the derivation of the weighting function. Most importantly, the above can only work if the data has a uniform mean over the domain, and a uniform variance; large scale spatial trends and heteroscedasticity will cause inaccurate results. These requirements are referred to as “second order stationarity”, or the intrinsic hypothesis in the literature. More formally expressed, the mean and variance

are translation invariant (subsetting the data at any point in the domain and taking the second order statistics will produce the same result everywhere).

These constraints can be relaxed somewhat if we Kriging only in a small moving neighbourhood (the assumption here is of local stationarity). However, in the face of strong non-stationarity, we simply reformulate as follows:

$$y = M .x + C$$

We simply replace the mean process c by a trend matrix C , the residuals of which display the statistical properties we require. Thus, the separation of the spatial trend into a large scale trend process (or drift), and a spatially correlated residual noise process is the rationale of non-stationary geostatistics.

The Kriging system is readily expanded to any number of dimensions, and 3D Kriging is not unusual in the literature. In our case we have two spatial dimensions (we are interested in the temperature experienced by the vegetative surface), plus time to deal with. It is important to explicitly model in space/time, as repeated spatial runs grossly overestimates the degrees of freedom, and, more practically, would cause continuity problems for simulations (weather fronts move in space and time).

Back to the problem at hand, the met data falls firmly into the non-stationary data camp. Thus the first task was to quantify and remove the large scale drift process. It is worth noting that the solution here is non-unique, and much effort has gone into removing only the simplest process which satisfies our requirements. It can be shown that allowable trend functions include the polynomial and exponential families, plus the Fourier series. All models used should be of low order (typically $k < 4$). Given that we are dealing with a time series, the Fourier model seems the most natural choice.

The model was fit to the data by first flattening the landscape via a lapse model of temperature against elevation ($p < 0.0001$), indicating a loss of around 3.5°C per Km rise. The data were then aggregated spatially by taking the median (regarded as the best measure of central tendency for skewed data). This robust ‘flat Earth average’ was then analysed by spectral methods (periodogram) to identify significant frequencies.

As it happens, we only require a single sinusoidal basis function to simulate the seasonal cycle, plus a linear warming term (0.39°C over 5 years), and the elevation lapse rate (Fig 2). The spatial trends are thus incorporated by modifying the intercept of the Fourier series. Thus we superimpose our temporal trend model onto the landscape (s subscript indicates spatial effect):

$$c = \alpha_s + \beta_1 \sin(\omega.t) + \beta_2 \cos(\omega.t)$$

where

$$\alpha_s = \alpha_0 + \alpha_1 t + \alpha_2 \text{elev}_s.$$

$$\omega = (2.\pi / 365)$$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
Alpha0	5.471	2.183 e-02	250.59	<2e-16
Alpha1	6.285 e-04	1.653 e-05	38.02	<2e-16
Alpha2	-3.568 e-03	1.694 e-05	210.68	<2e-16
Beta1	-2.489	1.221 e-02	203.87	<2e-16
Beta2	-5.069	1.208 e-02	419.49	<2e-16

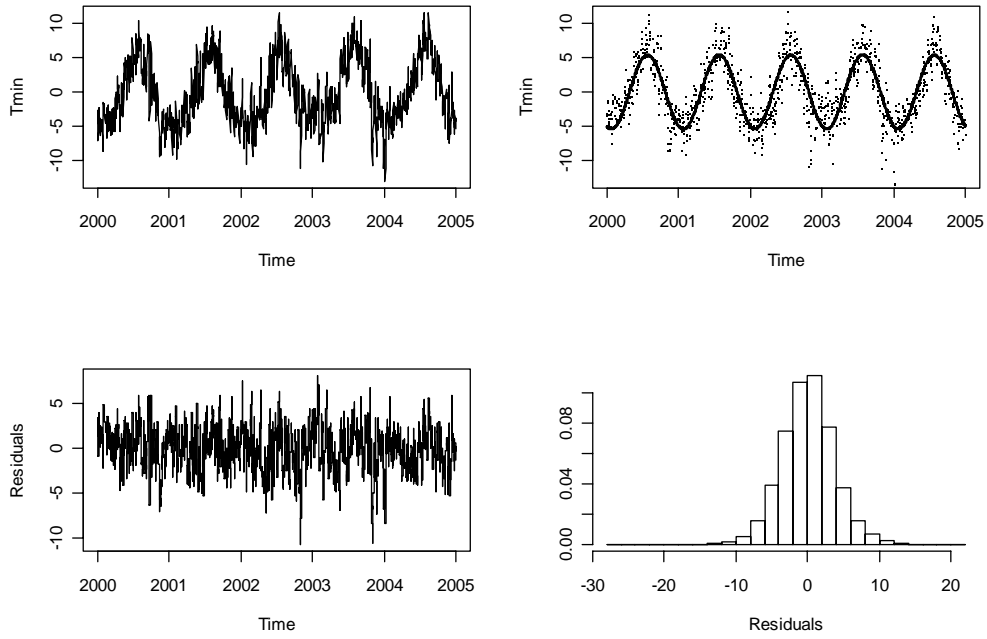


Fig 2: Aggregated time series, model, residual trace and histogram

All model effects are highly significant ($p < 0.0001$), and the r^2 is around 58%. The residual are a zero mean, (locally) second order stationary Gaussian process (Shapiro-Wilk normality test; $p < 0.7818$).

Having established a workable residual set, I then derived the temporal autocorrelation structure (Fig 3):

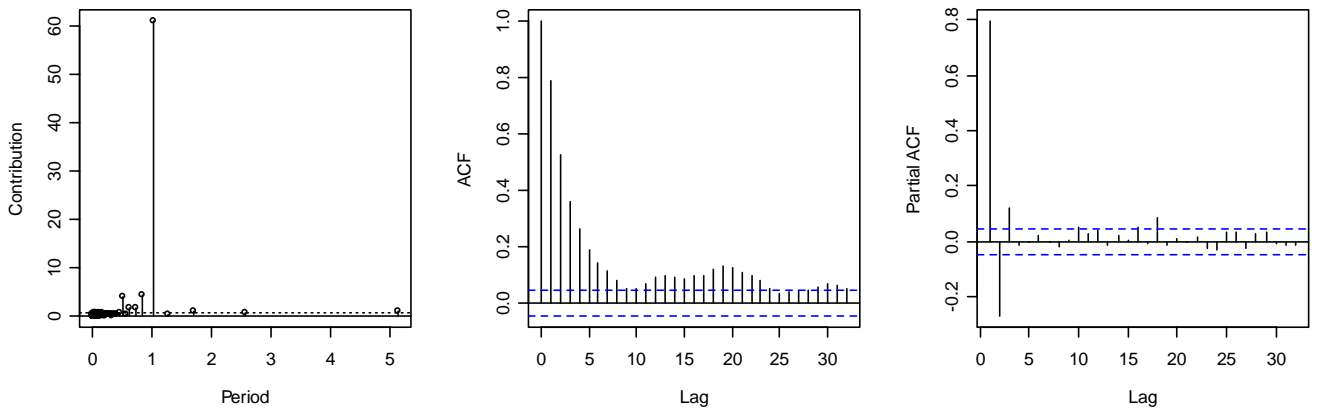


Fig 3: Periodogram of the temperature series, ACF of the model residuals and partial ACF of the residuals.

As expected, the dominant periodicity is the one year cycle, but there are interesting significant peaks at shorter and longer periods. These were tested as component of the trend model, but did not improve the AIC (Akaike's information criterion – a balance of improved model fit versus parsimony). It is likely that these are artefacts caused by spectrum leakage. The autocorrelation structure of the residuals is highly structured, and tails off below the significance cut-off within 30 days. The partials show few significant interactions after 5 days, with a marginally significant cycle at 18 days.

To assess the spatial trends, the data were disaggregated spatially, but averaged temporally by taking the mean of the residual series for each location (Fig 4).

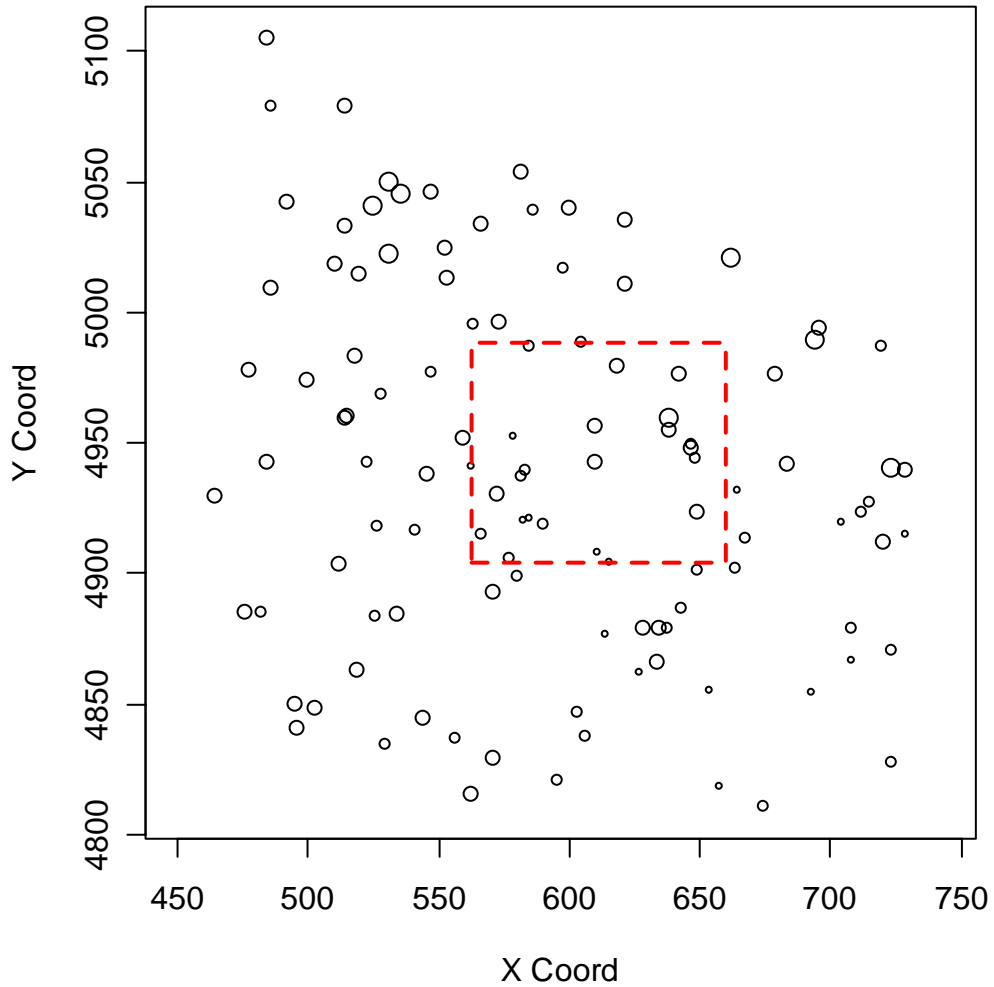


Fig 4: Data posting of 113 stations; point size is proportional to the data quartiles i.e. small = colder than the median. Broken line indicates area of interest.

A cursory inspection of the plot indicates lower values in the south east of the state, with a band of higher values to the north. From this data, an empirical variogram was constructed, which appears stable. Analysis of the variogram indicated an inflection at a separation of around 100km, where the data have a significantly higher variance than the global median (Fig 5).

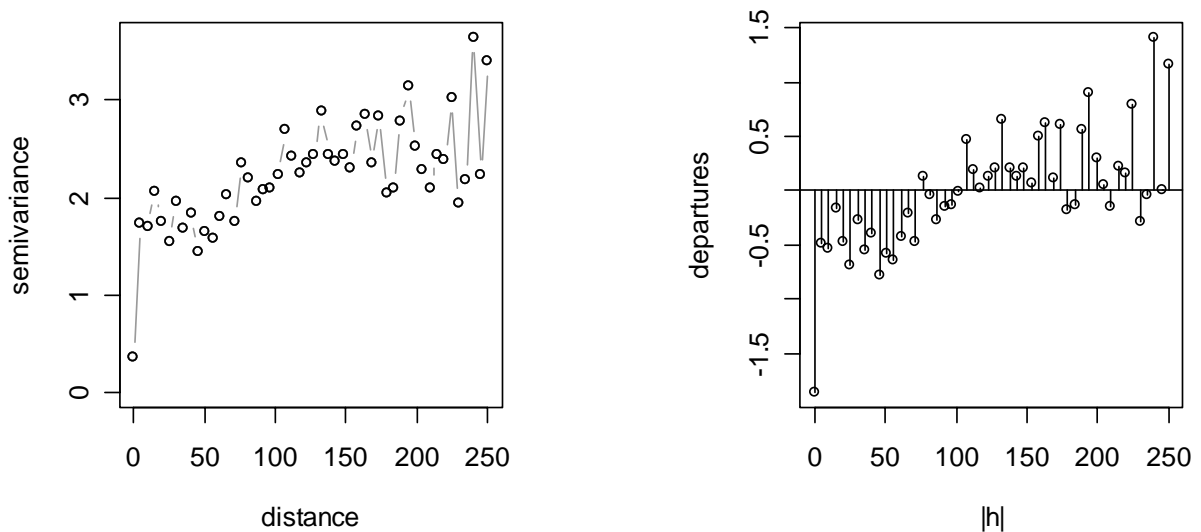


Fig 5: Empirical Semivariogram and departures from median; note the inflection in sign at 100 km separation.

A series of models were considered and fit to the empirical plot, initially by eye, then via a series of optimisation routines, using both weighted least squares (WLS), or maximum likelihood (ML) fits. Models were selected via cross validation with the data, with the criterion of minimising the RMS error. The selected models were:

1. Exponential model, REML (Reduced Error Maximum Likelihood) fit
2. Spherical model, ML fit
3. Cubic model (stable alternative to Gaussian), WLS fit with Cressie's robust weights.

As indicated, a variety of classical least squares approaches, and model based computational optimisation methods were employed. A comparison of model fits and Kriging outputs indicates that ranging from the first to the third; the models become smoother in their surface characteristics (Fig 6). The smoothest surface (cubic model) has the lowest RMS error, and is thus considered the best model. Fig 7 shows the selected model fit and maximum likelihood envelope (the max-min range of data values resulting from simulation via this model – $n=999$).

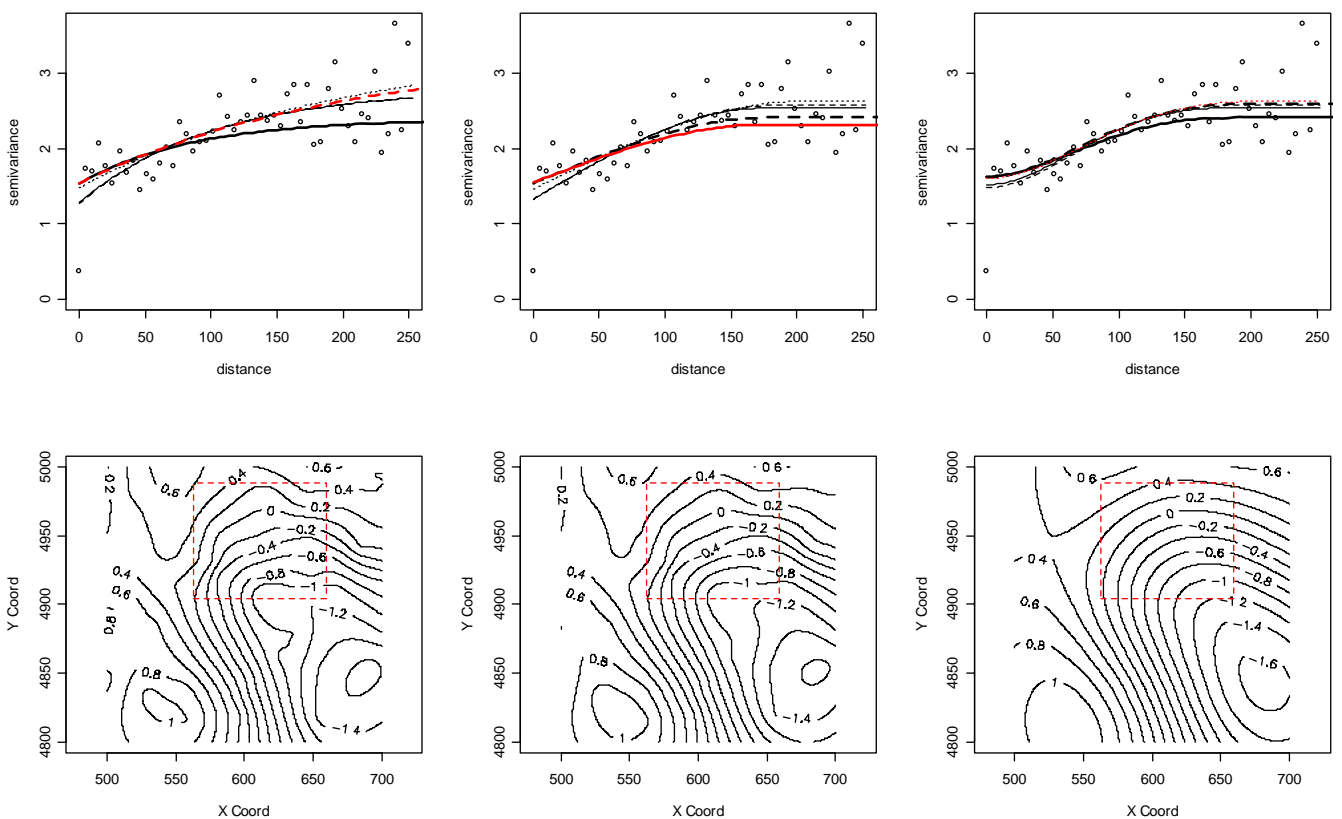


Fig 6: Kriging surfaces from the 3 selected models, with modelled semivariograms. In each case the thin solid line is the wls fit, the thin broken line is the ols fit, and the thin dashed line is the cressie fit. The heavy lines are the maximum likelihood estimates; the heavy solid line is the ML fit whilst the heavy broken line is the REML fit. The selected model is highlighted in red, as is the study area.

Note that the output of the Kriging is incredibly similar in each case, despite totally different functional forms, parameters and fitting techniques (even statistical paradigms come to that). This is actually unsurprising; it is known that the Kriging system is resistant to changes in the variogram model, but (in Noel Cressie's words) this does not mean we should not aim to make the most of our data.

To Do:

Currently I am working on some new code to take the average of the variograms for each time slice (i.e. model then average, rather than average then model); this should produce a more robust fit. I am able to

fit a fully partitioned model in one step now, by specifying the spatial trend with elevation in the variogram computation – this is advantageous when calculating variograms at each time step, as the elevation trend parameters are dynamic in time:

	Basal Temperature	Lapse Gradient
January	1.917752	-0.004061264
February	1.952599	-0.004170565
March	3.049877	-0.003866861
April	4.477266	-0.004054654
May	7.062704	-0.003792643
June	9.853485	-0.003466816
July	11.620311	-0.002307653
August	11.502209	-0.002421996
September	9.552992	-0.003219401
October	6.454181	-0.003857118
November	2.974849	-0.004028790
December	2.294643	-0.004037244

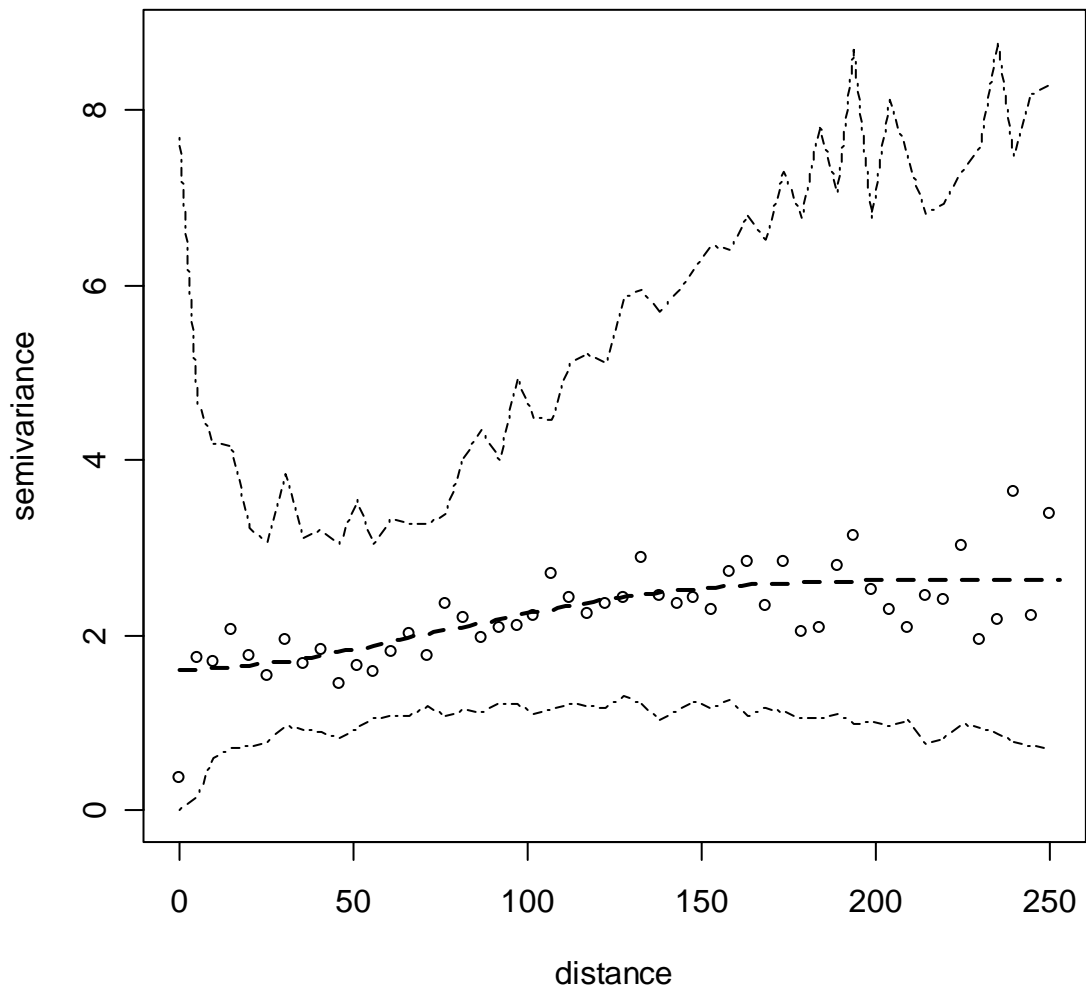


Fig 7: Best fitted variogram; cubic model by wls (cressie weights); model convex hull indicated by the light broken line.