Evaluating data interpolation in moving sparse noisy data to a uniform grid

Soumo Mukherjee*†, Daniel Caya†, René Laprise†

1. The Problem

The Canadian Regional Climate Model (CRCM) (Caya and Laprise, 1999) is routinely used to provide regional climate change projections. In order to assess the quality of CRCM, simulations run in the past must be compared with observational data.

However, observations contain error, and the observational network is distributed inhomogeneously. In Canada the observational network is densest towards the South (near population centres), whereas model output is homogeneously distributed on a 45km X 45km grid, leaving open the question, how can a fair comparison be made?

We propose to use a multi-variate, noisy-data interpolator to grid the observational network. However before doing so, the performance of the interpolator itself must be appropriately understood. Thus in the present experiment, we take CRCM screen temperature over the Quebec Region and choose noisy data subsets (simulating observational networks), trying to reproduce the original field using our interpolator.

2. Methodology

a. The model

ANUSPLIN, developed at the Australian National University makes use of thin-plate smoothing splines to minimize noise, thus creating smooth fields (for a more complete description, see Hutchinson, 1997). Clearly, maintaining smooth fields comes at the cost of preserving data-fidelity. Through minimization of the appropriate penalty function, ANUSPLIN finds the optimal balance between exact data interpolation (keeping loyal to the data, leaving rough fields) and regression (producing a smooth, less loyal field), objectively.

b. The data

Presented here (Figure 1) is screen temperature (ST) produced over the Quebec region from top left (70°W, 66°N) to bottom right (70°W, 40°N), on July 1st, 1978 using the CRCM. This dataset was thinned by leaving only every Nth row and column (termed NxN). Another subset (STN) corresponding to actual observational stations present in Canada at the time, was used. The full (1x1) dataset was also sullied with random (uncorrelated) noise at the 2, 5, 10, and 20% (of the maximum field range ~25°C) levels (corresponding to ~±0.5, 1, 2.5, 5°C). At these noise levels, the thinning procedure was then used to obtain noisy data subsets.

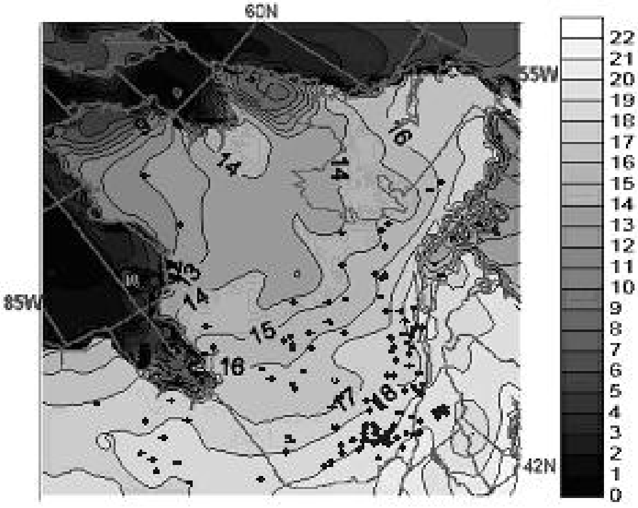

Figure 1. Summer Screen Temperature over Quebec with superimposed STN mask (black dots)

The experiment was then to interpolate from the NxN sparse fields to the original (dense) grid in order to determine the effect of selection (sparseness) as the fields deteriorate. Similarly the performance of ANUSPLIN was ascertained in the face of noise. Finally the noisy sparse sets were used to simulate an observational network where both effects were present.

c. Diagnostics

Principal diagnostics include difference maps depicting the difference between the fitted field and the original set, with the lighter colours depicting over-estimation and the darker, under-estimation.

The variance estimate, σ^2 (not shown), is the sum-of-squares of the fitted residuals, and represents the common data error. The model standard error, σ_m^2 (also not shown), represents the distributed Bayesian error-of-fit estimate. The prediction standard error, σ_p^2, represents the total error, as contributed to by both of these errors:

$$\sigma_p=(\sigma_m^2+\sigma^2)^{1/2} \qquad (1).$$

Hence this is the distributed error we could expect to calculate from our model given a certain data set.

3. Results

Once the original field is removed from the interpolated field with relatively high level of noise (10% or ±2.5°C) and sampled with few data points (4% of the original set), we can see that the interpolator performs well (Figure 2) with 60% of the difference over land within ±1°C (if we include water, this drops to 50% over the whole domain). Approaching the coastline, values are underestimated, as over-smoothing

***Corresponding author address:* Soumo Mukherjee, Sc. Terre-Atmosph., UQAM Salle PK-6435, B.P. 8888, Succ. "A", Montreal QC, Canada H3C 3P8.**
e-mail: soumo@sca.uqam.ca
†Université du Québec à Montréal

occurs in order to lessen the gradient. Reaching the water the opposite is true for the same reason. The effect of the land-sea interface is reduced in the St. Lawrence as it is interior to the domain. In the Hudson, Ungava, and James Bays, as well as Hudson Strait, and the Labrador Sea, there is less data on at least one side (domain border), so over-estimation occurs in accord with the rest of the domain (higher temperature).

The prediction standard error (equation 1) rises to ±3°C at noise level 10% and 5x5 data subset, but remains as low as just over ±1°C for no noise, 2x2 subsets, or using all data and noise levels less than 5% (not shown). Shown (Figure 3), are the σ_p for all cases with NxN subsets along the horizontal, and noise increasing vertically).

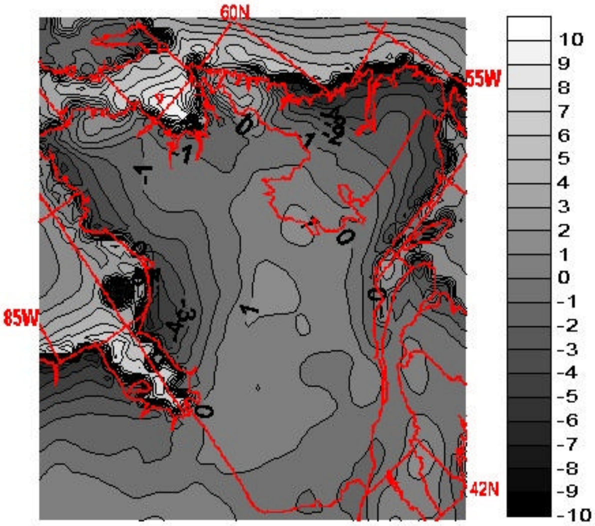

Figure 2. Difference of interpolated field (10% noise, 5x5 dataset) from original field (fig.1)

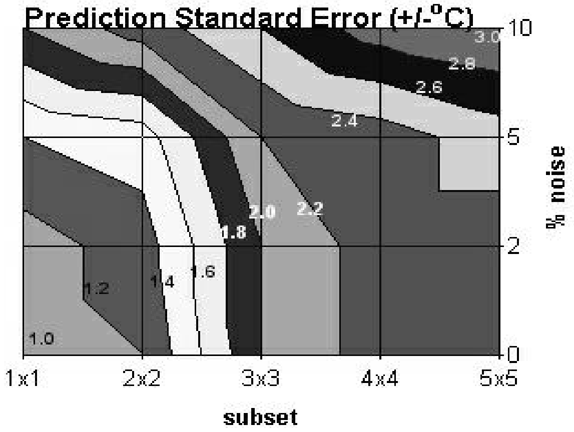

Figure 3. Prediction standard error composites

Interpolation from the realistic, scattered stations (see Figure 1 for distribution) is poor over water but good on land. The difference (interpolate less original field) map (Figure 4) shows a range that is greater by 2°C than that of the difference field of the 5x5 subset with 10% noise (Figure 2). The range is also shifted upwards (indicating overestimation) because the stations are mostly located on land, so values over water are under-represented. This is seen in over-estimated values upwards of 15°C in Hudson Bay. However, 57% of the total area is within ±1°C, due to better interpolation over land. Indeed, this is comparable to fields somewhere in the range of 2-5% noise for the 5x5 subset, or 5-10% noise for the denser 3x3 and 4x4 subsets.

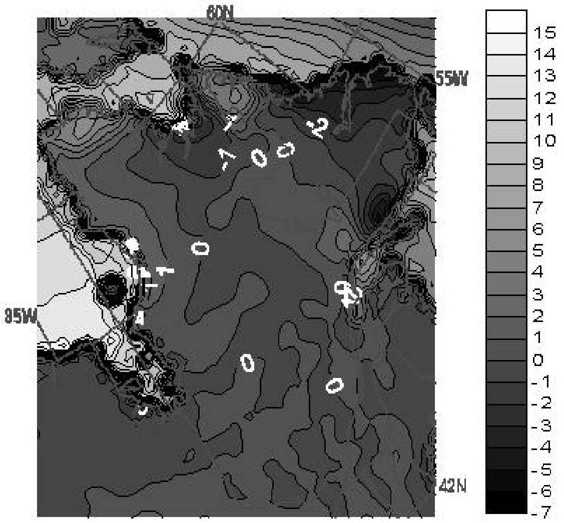

Figure 4. Difference of interpolated field (STN) from original

4. Conclusions

Problems occur mainly in regions of sharp change between land and water near the boundaries where the gradient can be as high as 8°C/100km. However, in the St. Lawrence for example, these steep (10°C/100km) gradients do not pose a problem as they are sufficiently interior (hence more anisotropically surrounded by data). Irregularly distributed data leads to exceedingly poor sampling in data-sparse regions, with better results over land.

Further work is being done to ascertain the effect of inclusion of the boundary, extra stations, as well as field continuity (precipitation, seasons, climatologies), and spatial distribution (non-homogeneous sets).

References

Caya, D. and Laprise, R. 1999: A semi-implicit semi-Lagrangian regional climate model: The Canadian RCM. *Mon. Wea. Rev.*, **127**, pp. 341-362.

Craven P. and Wahba G. 1979: Smoothing noisy data with spline functions. *Numerische Mathematik* **31**, pp. 377-403.

Hutchinson, M. F. 1997: ANUSPLIN VERSION 3.2, http://cres.anu.edu.au/outputs/anusplin.html.