Assessing uncertainty in numerical mesoscale weather prediction via ensembles of forecasts

Yulia Gel,
University of Washington, USA.

Joint work with Adrian Raftery and Tilmann Gneiting,
University of Washington, USA

Supported by the DOD Multidisciplinary University Research Initiative (MURI)

www.stat.washington.edu/tech.reports $\longrightarrow$ Technical Report no. 427

Criteria

- **Calibrated forecasts:**
  - if we define a probability interval, such as a 90$\%$ probability interval, then on average in the long run, 90$\%$ of such intervals contain the true value;
- **Sharp forecasts:**
  - if the distribution is more concentrated than the forecast distribution from climatology alone;

Outline

1. The basic idea
2. Data analysis
3. Direct method for simulations of forecast ensembles
4. Verification statistics
5. Circulant embedding method for simulation of gaussian random fields
6. Future development

$$\boxed{X \rightarrow M \rightarrow Y}$$

X are model inputs; Y are values to be forecast,

M is a mesoscale numerical prediction model (MM5)

(MM5 model is developed at the National Center for Atmospheric Research, Boulder, CO, USA, http://www.mmm.ucar.edu/mm5/mm5-home.html,)

Available Data

- **Current initial conditions from a big synoptic model on a grid, $\tilde{X}$.**
- **Historic information on $\tilde{X}$, also on a grid.**
- **Current MM5 forecasts on a grid, based on $\tilde{X}$, $\tilde{Y}$.**
- **Current direct observations, irregularly spaced, $\dot{X}$.**
- **Historic information about direct observations at stations, irregularly spaced, $\hat{X}$.**
- **Historic information on MM5 forecast $\tilde{Y}$, on a grid, $\hat{Y}$.**

The basic approach

Our main framework will be that of **Bayesian melding** (*Poole and Raftery, 2000, JASA*). This provides a way of combining up to 4 sources of information:

- prior information about inputs;
- prior information about outputs;
- likelihoods (data) about inputs;
- likelihoods (data) about outputs;

Basic idea: For now, focus only on a single predictive distribution of Y, $q(Y|\tilde{Y})$, and generate **statistical ensembles of forecasts** from $q(Y|\tilde{Y})$.

Alternative: Forme a single ”prior” distribution, $\pi(X|\tilde{X})$, and generate **statistical ensembles of initializations** from $\pi(X|\tilde{X})$.

Plan for the direct method

To define $q(Y|\tilde{Y})$ we use the model

$$\hat{X}(s,t) = a(s,t) + b(s,t)\hat{Y}(s,t) + \omega(s,t),$$

- s is a spatial coordinate; t denotes time;
- $a(s,t)$ and $b(s,t)$ are additive and multiplicative biases;
- $\omega(s,t)$ is a spatio-temporal stochastic process;
- $\omega(s,t) \sim \mathrm{MVN}\left(0, \Omega_\omega\right)$ with the covariance function of exponential form $\mathrm{Cov}(\omega(s_i), \omega(s_j)) = n + \sigma^2 e^{-d/r}$, where n is a nugget, σ^2 is a sill, r is a range.

For simplicity. *Let a and b be constants for now;*
$\hat{\mathbf{X}}(\mathbf{s},\mathbf{t}) \equiv \hat{\mathbf{X}}(\mathbf{s}), \tilde{Y}(s,t) \equiv \tilde{Y}(s), \omega(s,t) \equiv \omega(s)$

Then the predictive distribution of Y given $\tilde{Y}$ is defined

$$q(Y|\tilde{Y}) \sim N\left(a + b\tilde{Y}, \Omega_\omega\right)$$

where Ω_ω is the covariance matrix of $\omega(s)$.

Plan:

- Estimate a and b by linear regression.
- Estimate parameters of covariance functions of $\omega(s)$ by semivariogram model fitting.
- Simulate forecast ensembles from $q(Y|\tilde{Y})$ by circulant embedding method.

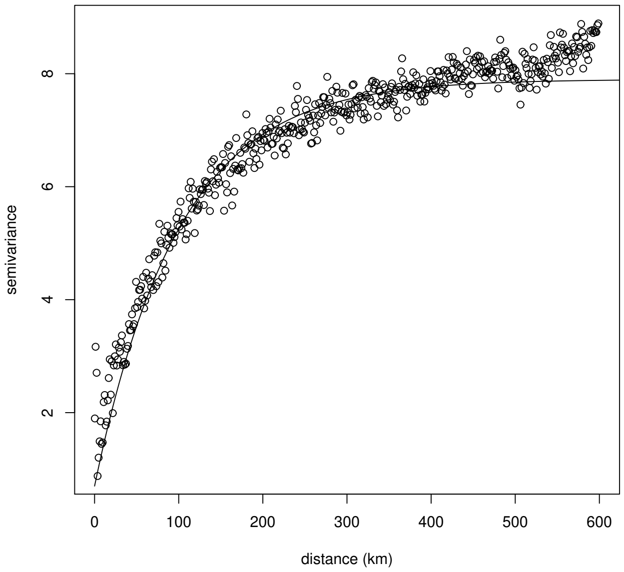
Temperature at surface level (T2).Variogram of $\omega(s)$.

Analysis is based on 102 days from January to June 2000

$$a = 1.31, \;\; b = 0.99$$

Semivariogram

$$V(s_1, s_2) = \frac{1}{2}\mathbf{E}(w(s_1) - w(s_2)) = \mathrm{cov}(0) - \mathrm{cov}||\mathrm{s}_1 - \mathrm{s}_2||$$



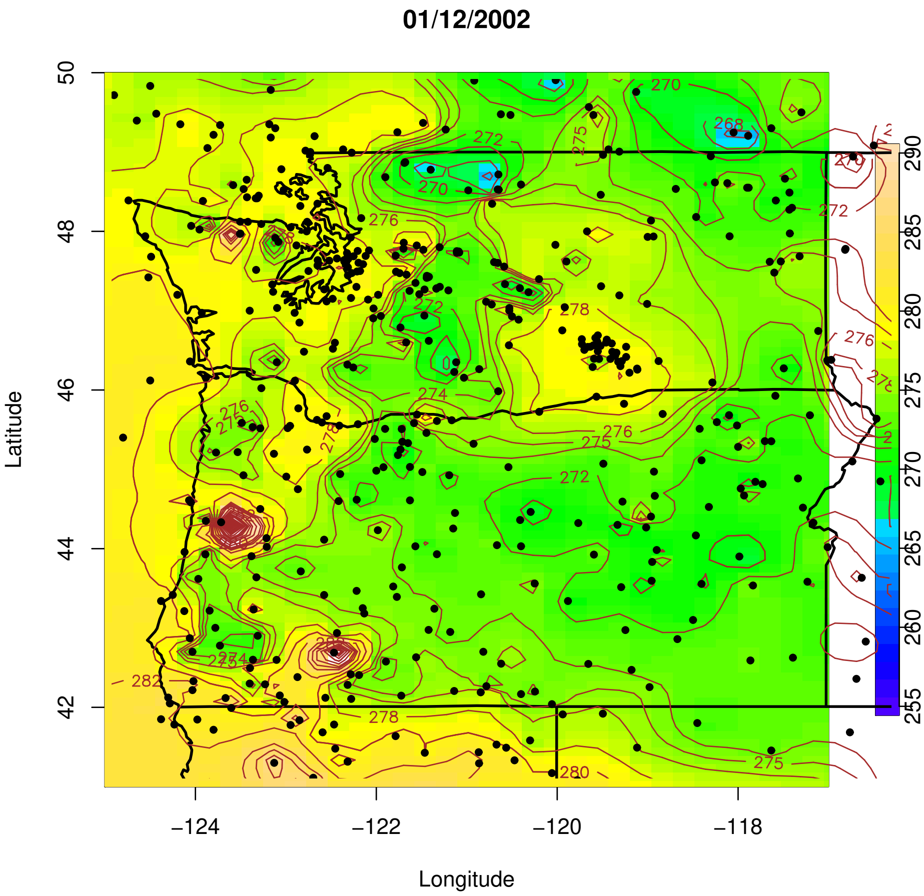
Covariance parameters for ω:

the range $\theta_\varepsilon = 0.0088,$

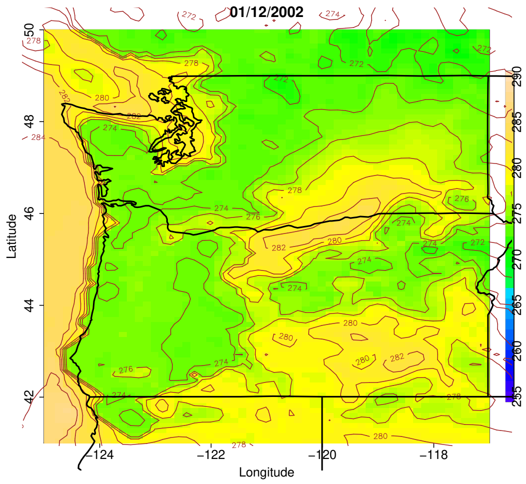
the sill $\sigma_\varepsilon^2 = 7.20,$

the nugget $= 0.51$.

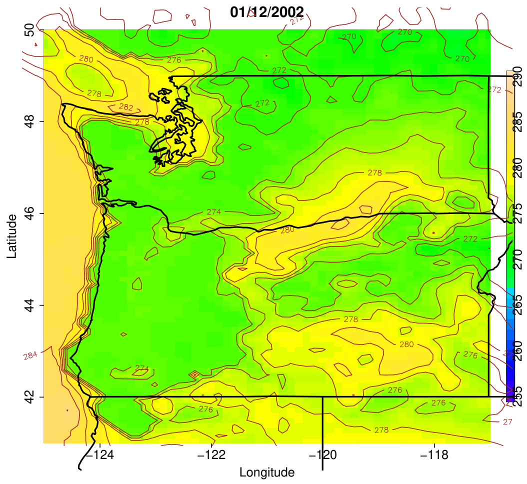
Observed temperature on Jan 12, 2002 at 0 hour GMT about 500 stations.



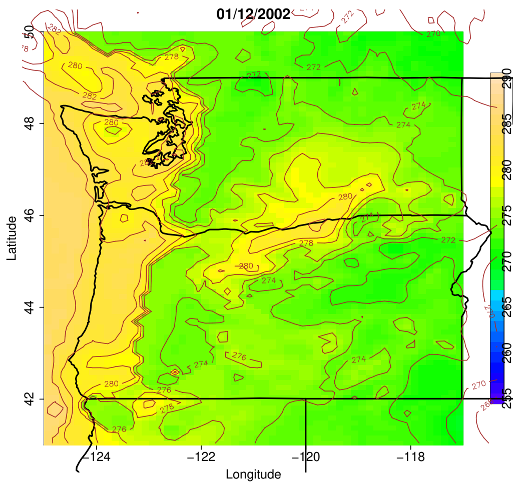
Temperature on January 12, 2002 at 0 hour GMT



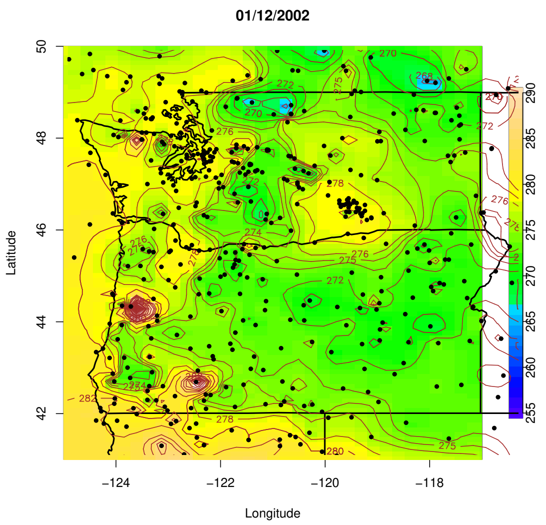
(a) The gridded MM5 forecast (init. on 10.01.2002 at 00GMT, 48 hours ahead), $\tilde{Y}$.



(b) The predictive mean, $a + b\tilde{Y}$.



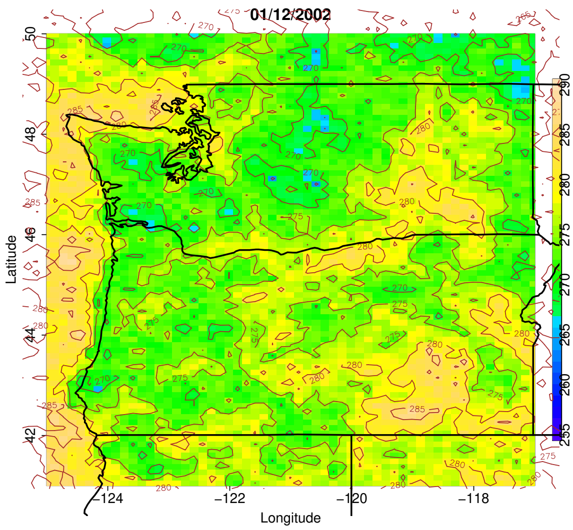
(c) Current estimation of T (the gridded MM5 output) on 12.01.2002 at 03GMT.

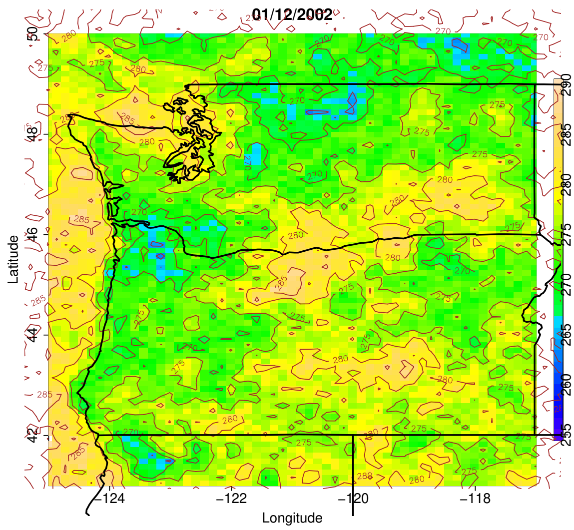


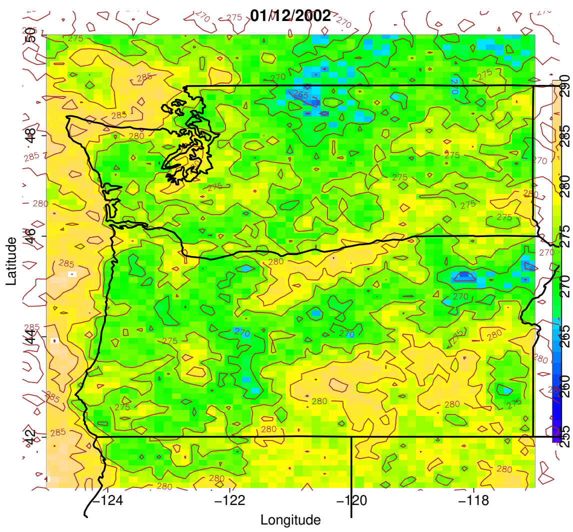
(d) Observed temperature at 500 stations) on 12.01.2002 at 00GMT

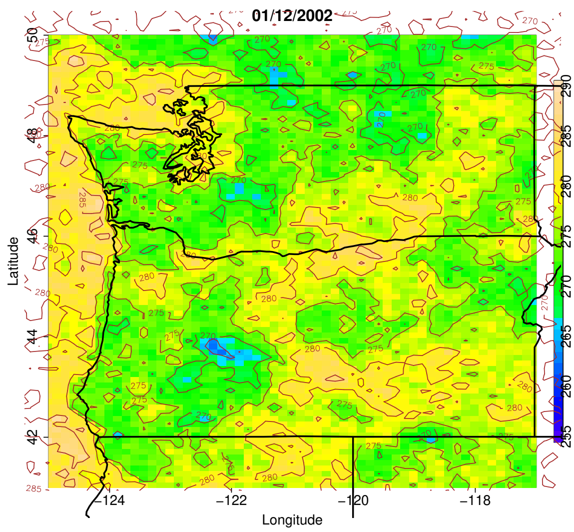
.

Simulated ensemble of forecasts on a grid. Temperature on January 12, 2002 at 0 hour GMT

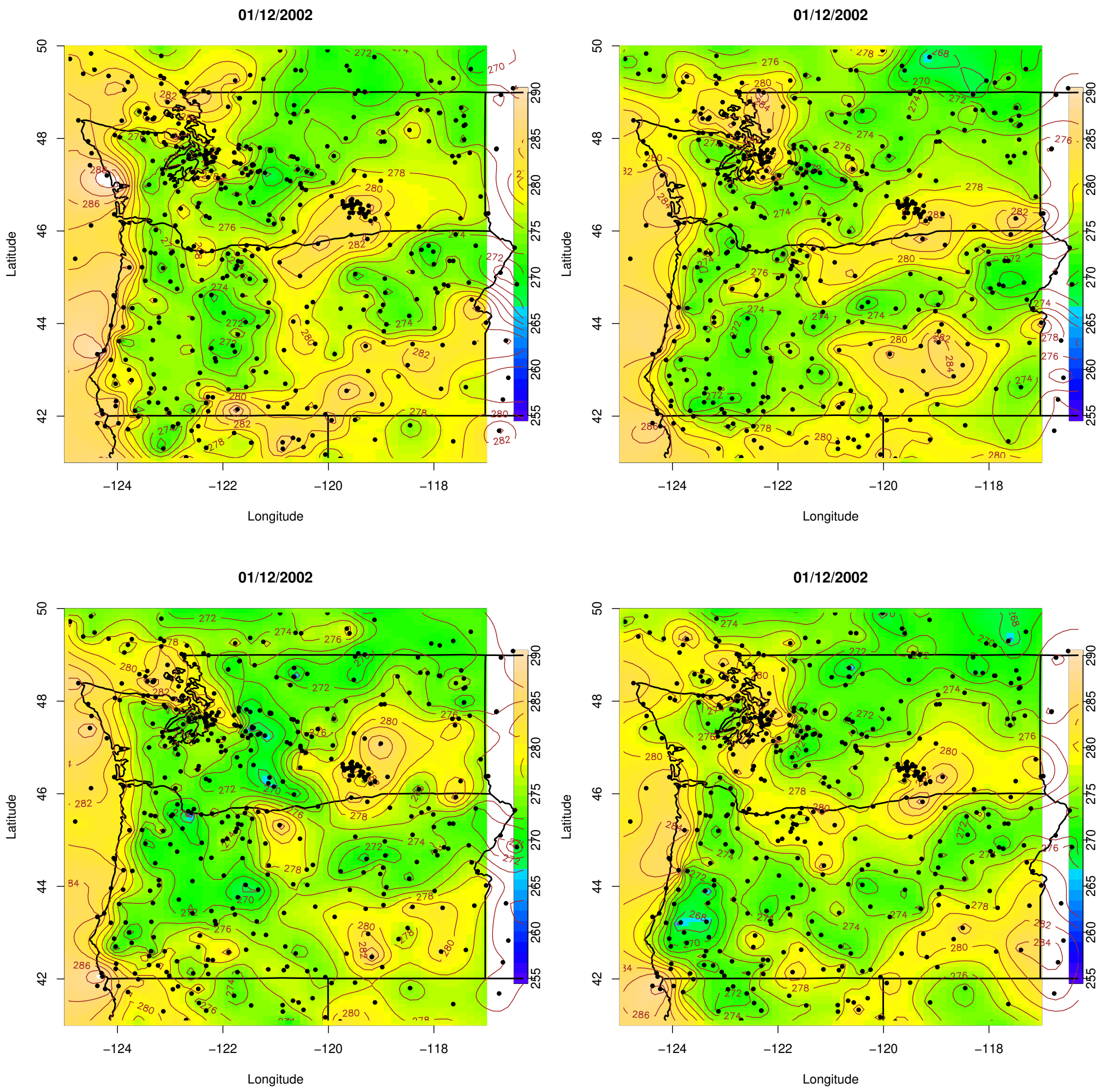








Map of the simulated forecasts interpolated by kriging from the station locations. January 12, 2002 at 0 hour GMT



Verification statistics

Table 1: Empirical coverage of 66.7% and 90% prediction intervals for temperature.

Prediction int.	Ensemble Forecast
66.7%	68.13%
90%	91.6%

- **The prediction intervals are well calibrated.**

Table 2: Length of the prediction intervals.

	mean		standard deviation
	observed values	forecast ensembles	
T2	28.3	9.43	0.76

Circulant embedding method

Required to obtain a realization of a gaussian random process X of the length n on the interval $[0, 1)$:
$X = \{X(0), X(\frac{1}{n}), \ldots, X(\frac{n-1}{n})\}$ **with covariance function γ, $\gamma(k, h) = \gamma(|k - h|)$.**

Hence its covariance matrix

$$\Sigma = \begin{pmatrix} \gamma(0) & \gamma(\frac{1}{n}) & \ldots & \gamma(\frac{n-1}{n}) \\ \gamma(\frac{1}{n}) & \gamma(0) & \ldots & \gamma(\frac{n-2}{n}) \\ \vdots & \vdots & & \vdots \\ \gamma(\frac{n-1}{n}) & \gamma(\frac{n-2}{n}) & \ldots & \gamma(0) \end{pmatrix}.$$

Let $m \geq 2(n-1)$. Embed Σ in a symmetric circulant matrix

$$C = \begin{pmatrix} c(0) & c(1) & \dots & c(m-1) \\ c(m-1) & c(0) & \dots & c(m-2) \\ \vdots & \vdots & & \vdots \\ c(1) & c(2) & \dots & c(0) \end{pmatrix},$$

with

$$\begin{aligned} c_j &= \gamma\left(\frac{j}{n}\right); \qquad 0 \leq j \leq m/2 \\ &= \gamma\left(\frac{m-j}{n}\right); \quad m/2 < j \leq m-1. \end{aligned}$$

Note that C is a covariance matrix of stationary stochastic process Y at m equally spaced points on a circle, $C = Q\Lambda Q^*$, where $\Lambda = diag\{\lambda_1, \lambda_2, ..., \lambda_m\}$, λ_i is an eigenvalue of C; Q is formed by the eigenvectors of C, $q_{jk} = m^{-1/2} exp\left(-\frac{2\pi ijk}{m}\right)$.

Then

$$Y \equiv Q\Lambda^{1/2}Q^*Z, \quad Z \sim N(0, I).$$

References:

- Grace Chan ”An Effective Method for Simulating Gaussian Random Fields”, *American Statistical Association Proceedings of the Statistical Computing Section, 133-138, 1999.*
- Andrew T.A. Wood, Grace Chan ”Simulation of Stationary Gaussian Processes in $[0,1]^d$”, *Journal of Computational and Graphical Statistics, 3, 409-432, 1994.*
- C. R. Dietrich, G. N. Newsam ”Fast and Exact Simulation of Stationary Gaussian Processes through Circulant Embedding of the Covariance Matrix”, *SIAM Journal on Scientific Computing , 18(4), 1088-1107, 1997.*
- C. R. Dietrich, G. N. Newsam ”Fast and Exact Method for Multidimensional Gaussian Stochastic Simulations”, *Water Resourses Research, 29(8), 2861-2869, 1993.*

Future directions

- **Analysis of spatio-temporal nonstationarity:**
  - *sill is a spatio-tempral stochastic process (e.g. via the Bayesian approach);*
  - *range is a spatio-tempral stochastic process (e.g. the norm trasformation and the Bayesian approach);*
  - *the kernel approach;*
  - *analysis of spatio-temporal separability, anisotropy analysis;*
- **Change of support (comparison of the range and smoothness parameters of the model output versus the observed data)**
- **Multivariate and multilevel statistical ensembles of forecasts based on multimodel initializations**