Grid-Based Bias Correction for Mesoscale Weather Forecasting: a statistical insight

Yulia Gel
University of Waterloo

Supported by the DOD Multidisciplinary University Research Initiative (MURI) and the National Sciences and Engineering Research Council of Canada

Bias=Forecast-Observation

Outline

- **Local Observation Method (LOB) –temporally weighed spatial composition of observations (Wedam, Mass and Steed (2005) and Tebaldi (2002)).**
- **The non-parametric regression method (CART-ACE) – elaboration of MOS:**
  - *spatial analysis of bias using non-parametric techniques (CART and ACE);*
  - *temporal analysis of bias;*
  - *combined spatio-temporal linear model.*
- **Combination of the previous two methods.**

Local Observation Method (LOB)

- Define the neighbor sites with available measurements of bias, e.g. stations with observations or grid points with available re-analysis.
- Construct a weighed function of the past biases observed at the neighbors during a certain time window T.
- Subtract the obtained function from the current forecast at the site of interest.

Questions:

- How to define the neighbors?
- How to define the optimal time window T?
- Should we weigh information differently for different neighbors and time points?

Sites are neighbors if:

- are located in a reasonable proximity;
- belong to the same terrain category;
- have similar elevation.

Wedam, Mass and Steed (2005) suggest to use only the dates with similar forecast, i.e. $\pm 5^\circ$, – regime adaptive ”Obs-Based” approach.

How far is to go geographically to search for neighbors?

Subjective Choice: Wedam, Mass and Steed (2005) suggest to expand geographically until find 5 neighbors.

Semivariogram Approach: estimate the spatial covariance and check at what distance it is nearly 0.

Aggregate the bias over certain time window which is sufficient enough to neglect fluctuations and to assume spatial (second order) stationarity and isotropy. Let $w(s)$ be the time aggregated bias at site s.

Then construct a non-parametric semivariogram,

$$\frac{1}{2}\,\text{Var}\left(w(s_1) - w(s_2)\right) = \text{Cov}(0) - \text{Cov}(||s_1 - s_2||),$$

which may be fitted for example with the parametric exponential covariance function

$$= \rho + \sigma^2 \left(1 - e^{-||s_1 - s_2|| r}\right)$$

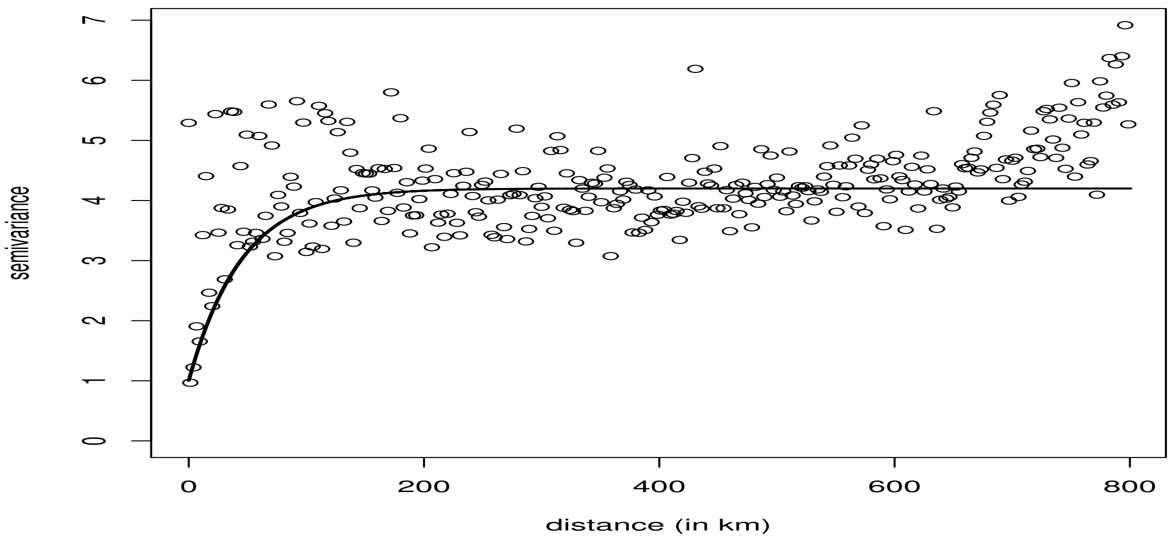
for $s_1 \neq s_2$ where $||\cdot||$ is the Euclidean norm.

In geostatistical terminology,

- ρ is the **nugget** effect–measurement error variance of observations,
- $\rho + \sigma^2$ is the **sill** – the marginal variance of $w(s)$,
- r, or **range**, is the distance at which the observations are (almost) uncorrelated.

Example. MM5 48 hours ahead forecast of surface temperature (T2) initialized at 00Z. Variogram of $\omega(s)$. 3 January 2001 to 31 December 2001.

Semivariogram $\frac{1}{2}$ **Var** $\left(w(s_1) - w(s_2)\right) = \rho + \sigma^2 \left(1 - e^{-||s_1 - s_2||r}\right)$



the range $r \simeq 120$

- The elevation criterion ± 200m is chosen empirically.
- Time window T is chosen as the best window T among $T = 1, 2, \ldots, 28$ past days.

Modifications:

- **Exponential weighting (EW) in time**, i.e. the information on t days back is weighed as e^{-t}.
- **Exclude outliers** in the neighbor data, i.e. use only the ”old” bias that fall into 95% range.

  The 95% range is estimated **robustly** over the recent biases in the neighborhood using **median absolute deviation, mean** and **quantiles of standard normal distribution.**

Figure 1: Mean absolute error (MAE) for various time windows T, i.e. $T = 1, \ldots, 28$ days, for the "Obs-Based", LOB and robust LOB (RLOB) methods with simple averaging (SA) and exponential weighting (EW) schemes.

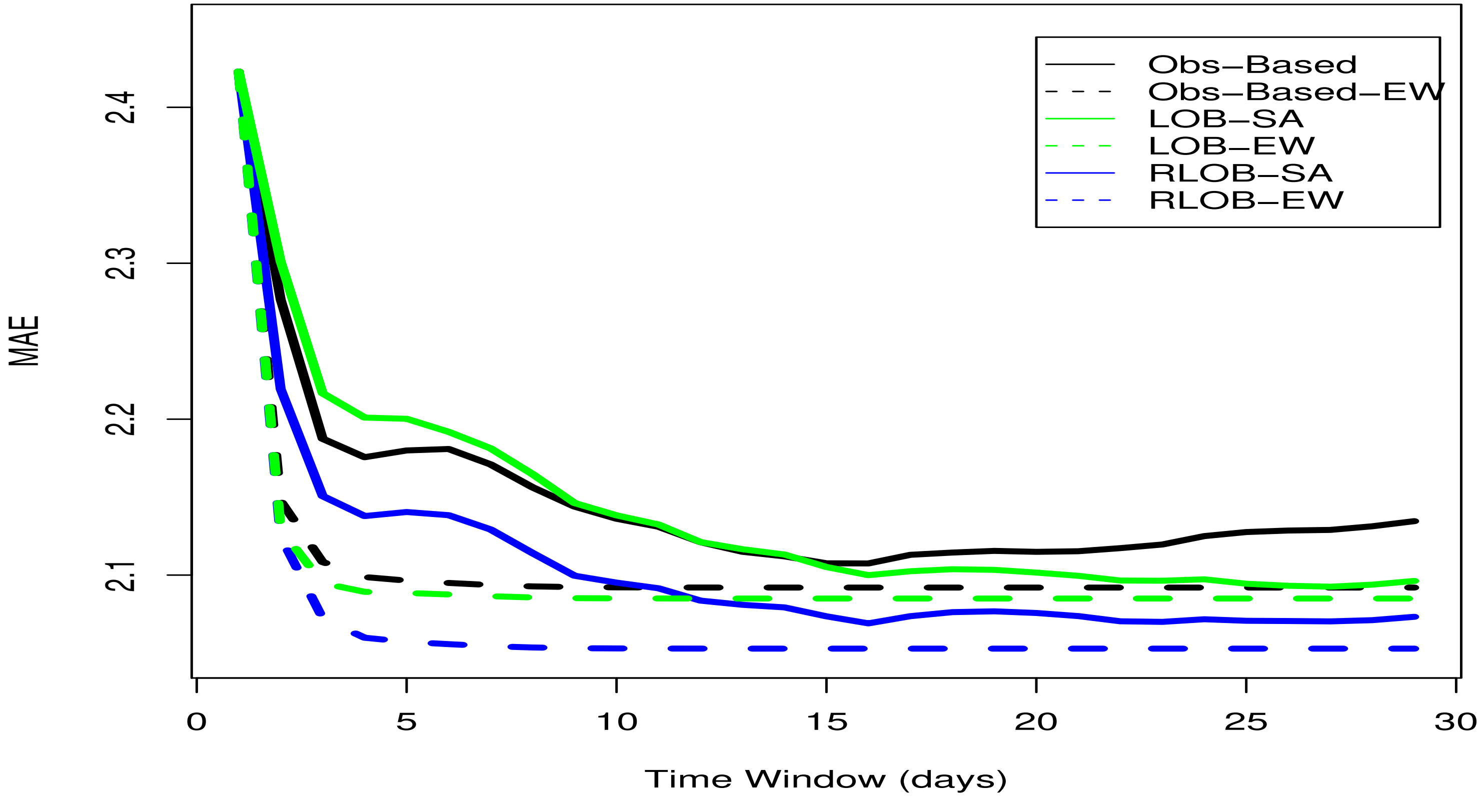


Table 1: MAE of MM5 forecasts before and after bias removal using the LOB, RLOB and ”Obs-Based” methods with simple averaging (SA) and exponential weighting (EW) schemes over the period from 3 January 2001 to 31 December 2001. The time window T is 14 days.

Raw	2.42
LOB-SA	2.10
RLOB-SA	2.07
”Obs-Based”	2.11
LOB-EW	2.08
RLOB-EW	2.05
”Obs-Based”-EW	2.09

Table 2: MAE of MM5 forecasts at 8 stations before and after bias removal using the RLOB and ”Obs-Based” methods with simple averaging (SA) and exponential weighting (EW) schemes over the period from 3 January 2001 to 31 December 2001. Time window is 2 weeks.

	Raw	RLOB-SA	”Obs-Based”	RLOB-EW	”Obs-Based”-EW
KEPH	2.85	1.92	2.09	1.96	2.06
KOLM	1.85	1.74	1.78	1.73	1.76
KGEG	2.25	2.09	1.87	1.86	1.81
KSEA	1.83	2.17	2.08	2.09	2.08
KBNO	2.71	2.71	2.71	2.71	2.71
KKLS	2.05	1.91	1.93	1.98	1.98
KYKM	3.01	2.05	2.24	2.07	2.23
KEUG	2.52	2.06	2.17	2.08	2.10
Average	2.73	2.08	2.11	2.06	2.09

Table 3: Mean and standard deviation of the biases of MM5 forecasts at 8 stations before and after bias removal using the RLOB and "Obs-Based" methods with simple averaging (SA) and exponential weighting (EW) schemes over the period from 3 January 2001 to 31 December 2001. Time window is 2 weeks.

	Raw	RLOB-SA	"Obs-Based"	RLOB-EW	"Obs-Based"-EW
KEPH	-2.22 (2.65)	-0.30 (2.47)	-1.00 (2.42)	-0.57 (2.42)	-0.82 (2.48)
KOLM	-0.70 (2.35)	0.17 (2.35)	-0.07 (2.37)	0.02 (2.29)	-0.07 (2.32)
KGEG	-0.75 (2.78)	1.01 (2.55)	0.31 (2.45)	0.73 (2.43)	0.44 (2.41)
KSEA	0.38 (2.42)	1.24 (2.59)	1.01 (2.58)	1.10 (2.53)	1.05 (2.56)
KBNO	-0.99 (3.30)	-0.99 (3.30)	-0.99 (3.30)	-1.00 (3.29)	-1.00 (3.29)
KKLS	-0.99 (2.48)	-0.31 (2.52)	-0.54 (2.47)	-0.41 (2.57)	-0.48 (2.51)
KYKM	-2.20 (2.93)	-0.28 (2.67)	-0.97 (2.68)	-0.58 (2.61)	-0.88 (2.70)
KEUG	-1.10 (2.94)	0.26 (2.68)	-0.20 (2.77)	0.03 (2.69)	-0.18 (2.71)
Average	-1.07 (2.73)	0.1 (2.64)	-0.31 (2.63)	-0.08 (2.61)	-0.24 (2.62)

Table 4: The ratio of improve and hurt cases (in %) for more than 2C for the RLOB and ”Obs-Based” methods with simple averaging (SA) and exponential weighting (EW) schemes over the period from 3 January 2001 to 31 December 2001. Time window is 2 weeks.

	RLOB-SA	”Obs-Based”	RLOB-EW	”Obs-Based”-EW
KEPH	26.4/4.8	16.4/1.1	29.1/5.6	23.6/5.4
KOLM	0/0	0/0.3	3.7/1.7	4.3/2.8
KGEG	14.9/12.7	13.5/3.1	20.3/7.6	16.6/5.4
KSEA	0/3.1	0.3/2.0	1.1/6.8	1.4/7.9
KBNO	0.3/0	0/0	0/0	0/0
KKLS	0/0	0/0	1.4/2.2	2.5/2.0
KYKM	29.7/5.4	17.0/1.4	30.3/4.0	25.5/5.7
KEUG	20.7/5.0	12.0/2.8	17.4/5.6	14.8/3.6
Average Ratio	$\frac{91.6}{31.0} = 2.95$	$\frac{59.3}{10.7} = 5.54$	$\frac{103.2}{33.5} = 3.08$	$\frac{88.8}{32.8} = 2.71$

The CART-ACE method – elaboration of MOS

Goal is to construct a linear spatio-temporal model which approximates non-linearies in the bias

Potential Predictors:

- Longitude, lon.
- Latitude, lat.
- Altitude, alt.
- Piece-Linear Functions from longitude, latitude and altitude.
- Land use categories, lu.
- sin/cos of Julian days of the year.

Spatial analysis of bias

- The bias at each station location is aggregated in time
- We apply Alternating Conditional Expectation (ACE) regression to the temporally aggregated bias and obtain k new linear predictors.
- We apply Regression Trees (CART) to the temporally aggregated bias and obtain l new linear predictors.
- We combine $k + l$ predictors from ACE and CART and m categorical terrain predictors and select n the most important predictors by the Akaike Information Criterion (AIC).

Figure 2: ACE transformation of longitude (lon), latitude (lat) and altitude (alt) (based on 100 stations and first 90 days in 2001).

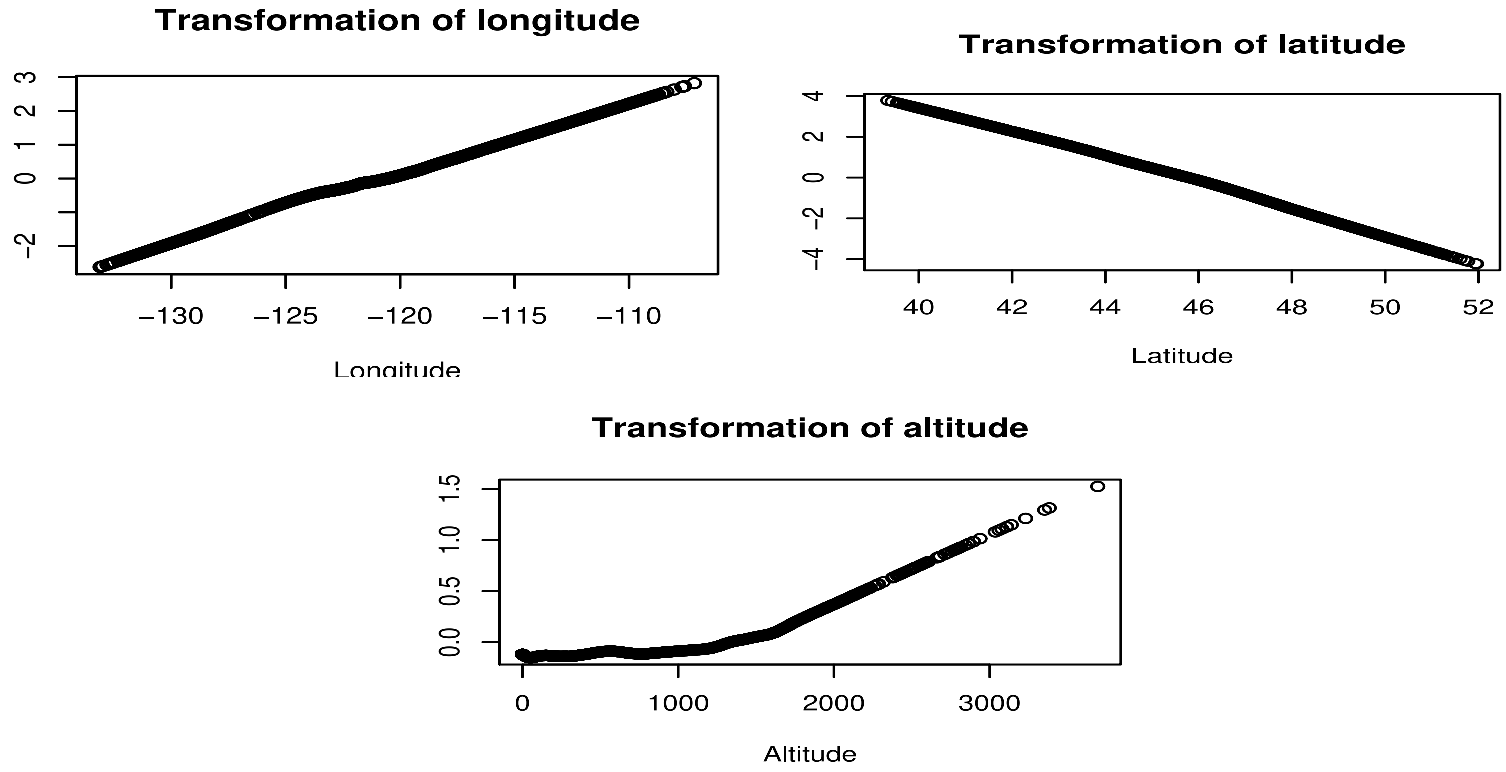
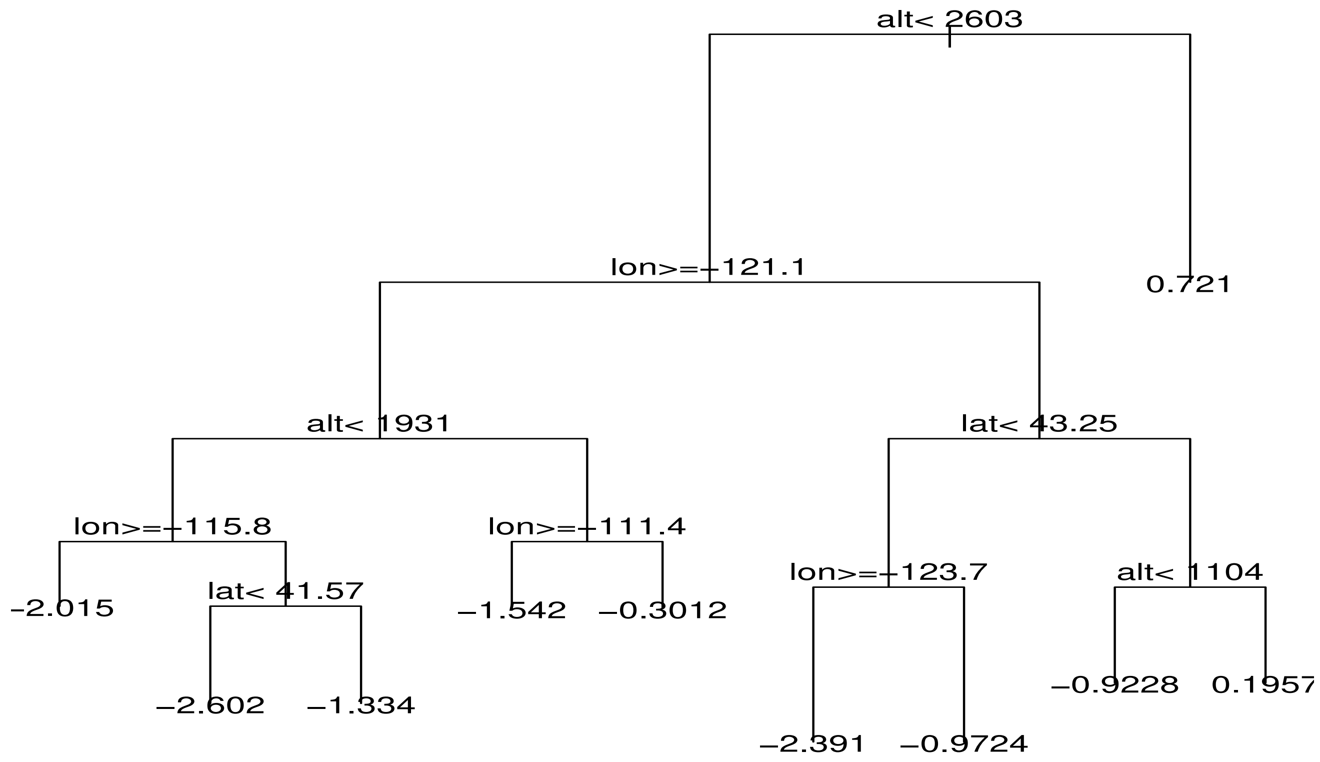


Figure 3: Regression tree of mean error vs. longitude (lon), latitude (lat) and altitude (alt) (based on 100 stations and first 90 days in 2001).



Using the AIC we obtain 8 main spatial binary predictors:

$$
\begin{aligned}
z_1 &= \mathrm{alt}_{[\mathrm{if\ alt}\geq 1000]}, \\
z_2 &= \mathrm{lon}_{[\mathrm{if\ lon}\geq -124]}, \\
z_3 &= \mathrm{lon}_{[\mathrm{if\ lon}< -124]}, \\
z_4 &= 1_{[\mathrm{if\ alt}\leq 1931 \mathrm{\ and\ lon}\geq -115.8]}, \\
z_5 &= 1_{[\mathrm{if\ alt}\leq 1931, \mathrm{lon}<(-115.8) \mathrm{\ and\ lat}<41.57]}, \\
z_6 &= 1_{[\mathrm{if\ alt}\leq 1931, \mathrm{lon}<(-115.8), \mathrm{\ and\ lat}\geq 41.57]}, \\
z_7 &= 1_{[\mathrm{if\ alt}<2603, \mathrm{lat}<43.25 \mathrm{\ and\ lon}<(-123.7)]}, \\
z_8 &= 1_{[\mathrm{if\ alt}\geq 2603, \mathrm{lat}\geq 43.25]},
\end{aligned}
$$

z_1, z_2 and z_3 correspond to the ACE; $z_4 - z_8$ correspond to CART

Now we add binary predictors for the land use and circular temporal predictors:

$$\begin{aligned}\cos_k(d) &= \cos(2\pi kd/365) \\ \sin_k(d) &= \sin(2\pi kd/365)\,, \qquad k = 1, 2, 3, 4\end{aligned}$$

where d is the Julian day of the year.

The combination of all binary spatial, circular temporal and binary terrain predictors yields a linear regression model (CART-ACE model).

In addition, we can naturally incorporate the robust LOB method as an extra predictor into the CART-ACE model.

Table 5: MAE of MM5 forecasts before and after bias removal using the CART-ACE method and the combined LOB-CART-ACE method over the period from 3 January 2001 to 31 December 2001. The time window T is 14 days.

Raw	2.42
CART-ACE	2.03
RLOB-CART-ACE	2.02
LOB-SA	2.10
RLOB-SA	2.07
”Obs-Based”	2.11
LOB-EW	2.08
RLOB-EW	2.05
”Obs-Based”-EW	2.09

Table 6: MAE of MM5 forecasts at 8 stations before and after bias removal using the CART-ACE and LOB-CART-ACE methods from 3 January 2001 to 31 December 2001. Time window is 2 weeks. The EW scheme.

	Raw	CART-ACE	RLOB-CART-ACE
KEPH	2.85	1.93	1.80
KOLM	1.85	1.78	1.82
KGEG	2.25	1.95	2.00
KSEA	1.83	2.35	2.46
KBNO	2.71	2.19	2.19
KKLS	2.05	2.09	2.07
KYKM	3.01	2.10	1.94
KEUG	2.52	2.04	2.02
Average	2.73	2.05	2.04

Table 7: Mean and standard deviation of the biases of MM5 forecasts at 8 stations before and after bias removal using the CART-ACE and LOB-CART-ACE methods from 3 January 2001 to 31 December 2001. Time window is 2 weeks. The EW scheme.

	Raw	CART-ACE	RLOB-CART-ACE
KEPH	-2.22 (2.65)	-0.74 (2.40)	-0.33 (2.34)
KOLM	-0.70 (2.35)	0.30 (2.36)	0.48 (2.36)
KGEG	-0.75 (2.78)	0.73 (2.53)	1.09 (2.46)
KSEA	0.38 (2.42)	1.18 (2.80)	1.36 (2.83)
KBNO	-0.99 (3.30)	0.21 (2.93)	0.21 (2.93)
KKLS	-0.99 (2.48)	0.43 (2.68)	0.57 (2.70)
KYKM	-2.20 (2.93)	-0.71 (2.60)	-0.31 (2.53)
KEUG	-1.10 (2.94)	0.36 (2.62)	0.64 (2.58)
Average	-1.07 (2.73)	0.22 (2.62)	0.46 (2.59)

Table 8: The ratio of improve and hurt cases (in %) for more than 2C for the CART-ACE and LOB-CART-ACE methods over the period from 3 January 2001 to 31 December 2001. Time window is 2 weeks. The EW scheme.

	CART-ACE	RLOB-CART-ACE
KEPH	20.8/0.8	32.5/3.4
KOLM	4.8/5.1	8.8/7.7
KGEG	13.2/4.8	18.3/10.7
KSEA	1.7/10.4	2.3/13.8
KBNO	15.6/3.1	15.6/3.1
KKLS	6.7/10.4	9.8/11.8
KYKM	2.8/2.3	35.4/3.7
KEUG	19.3/4.8	21.3/6.2
Average Ratio	$\frac{10.61}{5.21} = 2.04$	$\frac{18.0}{7.55} = 2.38$

Conclusions for the CART-ACE method

- The highest MAE reduction is shown by the RLOB-CART-ACE method. However, the main advantage has been shown at ”KBNO” station that does not have neighbors.
- The reduction of the bias mean (should be 0) is worse for the CART-ACE method. The improve-and-hurt statistics is also worse for the CART-ACE method.
- The CART-ACE method is not fully automatic. It makes sense to use it for the sites with no neighbors or relatively small locally run projects.

Conclusions for the LOB method

- The robustification is a valuable assset. The best MAE and mean reduction is shown by the RLOB-EW method.
- The Exponential Weighting (EW) is essential. All neigbor-based methods show better results with the EW scheme.
- The richer spatial representation with the neighborhood defined by the semivariogram works equally or better in terms of the MAE reduction and the correction of the bias mean than the regime-adaptive ”Obs-Based” method of Wedam, Mass and Steed (2005).
- It makes sense to combine the RLOB-EW with the regime-adaptive ”Obs-Based” method of Wedam, Mass and Steed (2005).
- The LOB may be fully automatic!