Bayesian hierarchical modelling for data assimilation of past observations and numerical model forecasts

Stan Yip

Exeter Climate Systems, University of Exeter
c.y.yip@ex.ac.uk
Joint work with Sujit Sahu in University of Southampton

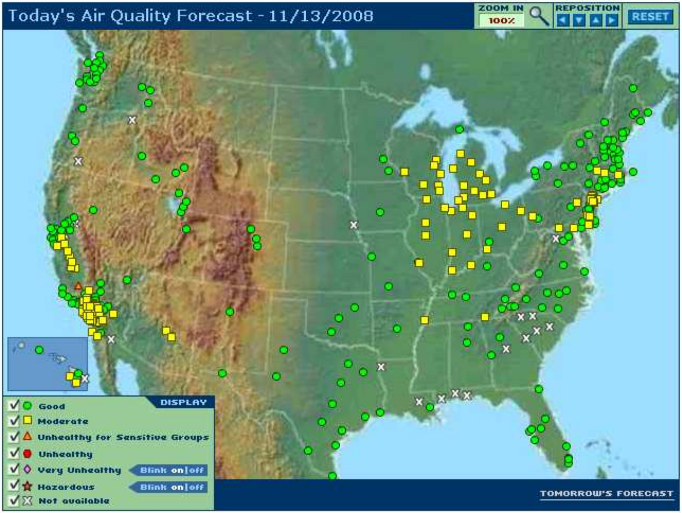
MPI-PKS, Dresden, 31st July 2009

Motivation

- Fusing ground level ozone concentration observations with computer deterministic model output.
- Improving biased forecast.
- Capturing spatio-temporal variation.
- Quantifying uncertainty through Bayesian probabilistic forecast.
- Producing high resolution maps.

Forecast

EPA's www.airnow.gov website

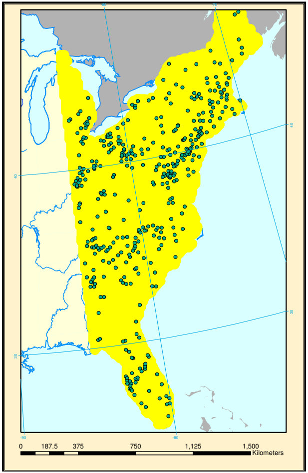


Ground Level Ozone

- Ground level ozone: bad health effects: primarily respiratory, lung function, coughing, throat irritation, congestion, bronchitis, emphysema, asthma.
- Ozone is a secondary pollutant.
- VOC's (Volatile Organic Compounds) - organic gases but really "chemicals that participate in the formation of ozone."
- Sunlight + VOC + NOx = Ozone.
- Meteorological conditions - sunlight, high temperature (so primarily from April to September), wind direction and wind speed. High spatial-temporal correlation.

Observations

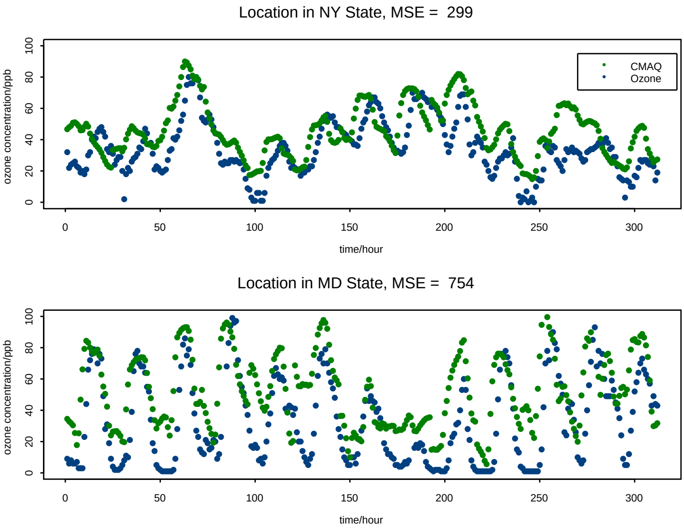
- 409 spatial point locations are in the area.
- Recorded hourly.
- Measured by unattended photometers.
- About 20 percent data is missing over 15 days.
- Sparse data.



CMAQ modelling system - Computer model output

- National Oceanic and Atmospheric Administration (NOAA) have designed the Community Multi-scale Air Quality (CMAQ) modelling system.
- The model is used by Environmental Protection Agency (EPA)
- CMAQ consists of a set of deterministic physical models from first principle.
- The forecasts are biased.
- Computer model outputs are in grid cell, but in the real siuation, we want point location prediction.
- Uncertainty has not been taken into account.

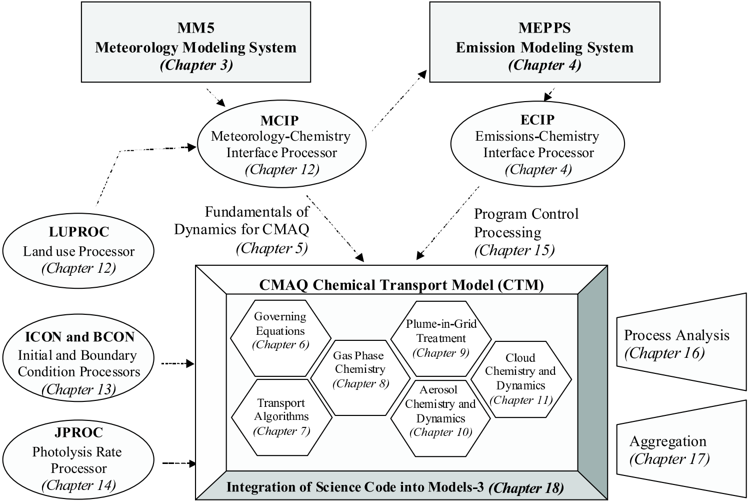
CMAQ



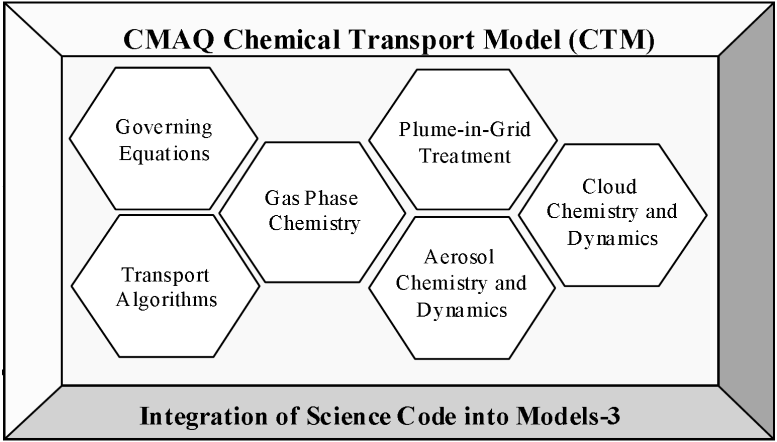
Location in NY State, MSE = 299
CMAQ
Ozone
ozone concentration/ppb
0
20
40
60
80
100
0
50
100
150
200
250
300
time/hour
Location in MD State, MSE = 754
ozone concentration/ppb
0
20
40
60
80
100
0
50
100
150
200
250
300
time/hour

CMAQ modules

Ching and Byun,1999



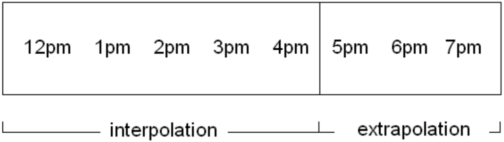
CMAQ modules



Problem

Daily 8-hour maximum Prediction

- 8-hour average ozone concentration is an important indicator for environmental monitoring.
- Measuring the daily 8-hour average maximum ozone concentration is required by the law.
- One day ahead 8-hour average maximum ozone concentration at an arbitrary location is needed.
- High resolution map can be produced from the prediction outputs.
- Obtaining forecasts within few hours.



Work done by others

- Fuentes and Raftery (2005) combine the computer model and observation by joint multivariate normal distribution.
- Zimmerman and Holland (2005) use different data sources with different measurement error and bias.
- Jun and Stein (2004) compare the correlation structure of computer model and observations.
- None of them deal with space-time forecast at the same time.
- The measurement is not ground truth.

Why do we adopt Bayesian approach?

- Probabilistic forecast addresses the uncertainty through distribution (pdf).
- Modelling becomes more flexible.
- Linear regression model doesn't work here, it cannot capture spatial correlation.
- The approach distinguish "ground truth", "measurement" and "biased forecast".

Model Structure

$$
\begin{tikzcd}
 & & \text{Historical Data} & \text{Forecasts} \\
\text{Observation} & & Z(\mathbf{s},t) & \\
\text{Ground Truth} & O(\mathbf{s},t-1) \arrow[r] & O(\mathbf{s},t) \arrow[u, "\epsilon(\mathbf{s},t)"'] \arrow[r] & O(\mathbf{s},t+1) \\
\text{CMAQ} & & x(\mathbf{s},t) \arrow[u] & x(\mathbf{s},t+1) \arrow[u]
\end{tikzcd}
$$

Measurement Equation: $Z(\mathbf{s}_i,t) = O(\mathbf{s}_i,t) + \epsilon(\mathbf{s}_i,t)$

System Equation: $O(\mathbf{s}_i,t) = \xi_t + \rho\, O(\mathbf{s}_i,t-1) + \beta_0\, x(\mathbf{s}_i,t) + \eta(\mathbf{s}_i,t)$

Model Specification

$$\begin{tikzcd}
 & & \text{Historical Data} & \text{Forecasts} \\
\text{Observation} & & Z(\mathbf{s},t) & \\
\text{Ground Truth} & O(\mathbf{s},t-1) \arrow[r] & O(\mathbf{s},t) \arrow[u, "\epsilon(\mathbf{s},t)"'] \arrow[r] & O(\mathbf{s},t+1) \\
\text{CMAQ} & & x(\mathbf{s},t) \arrow[u] & x(\mathbf{s},t+1) \arrow[u]
\end{tikzcd}$$

Measurement Equation: $Z(\mathbf{s}_i,t) \sim N(O(\mathbf{s}_i,t),\sigma_\epsilon^2)$,

System Equation: $\boldsymbol{O}(t) \sim N(\xi_t + \rho\,\boldsymbol{O}(t-1) + \beta_0\,\boldsymbol{x}(t), \sigma_\omega^2\Sigma)$,
where $\boldsymbol{O}(t) = (O(\mathbf{s_1},t),\ldots,O(\mathbf{s_n},t))'$,
$\boldsymbol{x}(t) = (x(\mathbf{s_1},t),\ldots,x(\mathbf{s_n},t))'$.

How do we forecast?

Posterior Predictive Distribution

The posterior predictive distribution of $Z(\mathbf{s}', t')$ is obtained by integrating over the unknown quantities with respect to the joint posterior distribution, i.e.,

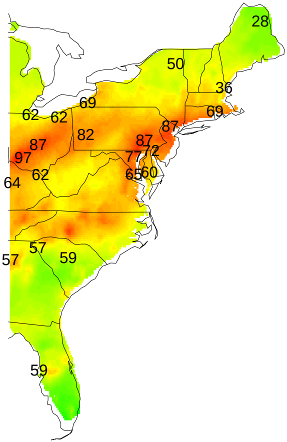
$$\begin{aligned} \pi\left(Z(\mathbf{s}', t')|\mathbf{z}\right) \quad = \quad & \int \pi\left(Z(\mathbf{s}', t')|O(\mathbf{s}', [t']), \sigma_\epsilon^2\right) \\ & \pi\left(O(\mathbf{s}', [t'])|\boldsymbol{\theta}, \mathbf{w}\right) \\ & dO(\mathbf{s}', [t'])\, d\boldsymbol{\theta}\, d\mathbf{w}. \end{aligned}$$

It can be done by Monte Carlo integration in the Markov chain Monte Carlo routine.

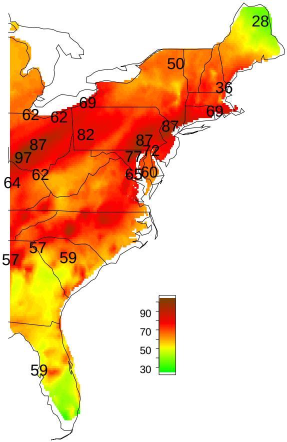
Prediction Maps

The 1-day ahead forecast surfaces on 11th Aug: Bayes and CMAQ

Bayes forecast map for the following day: 11th Aug

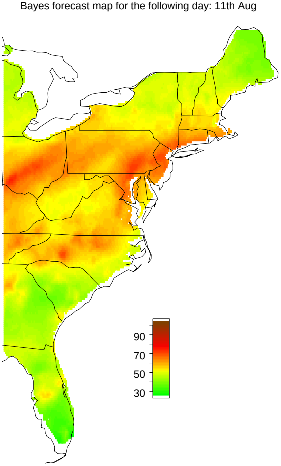


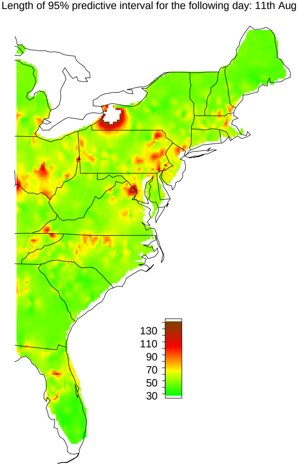
CMAQ forecast map for the following day: 11th Aug



Prediction Maps

The 1-day ahead forecast surfaces on 11th Aug: Bayes and its uncertainty





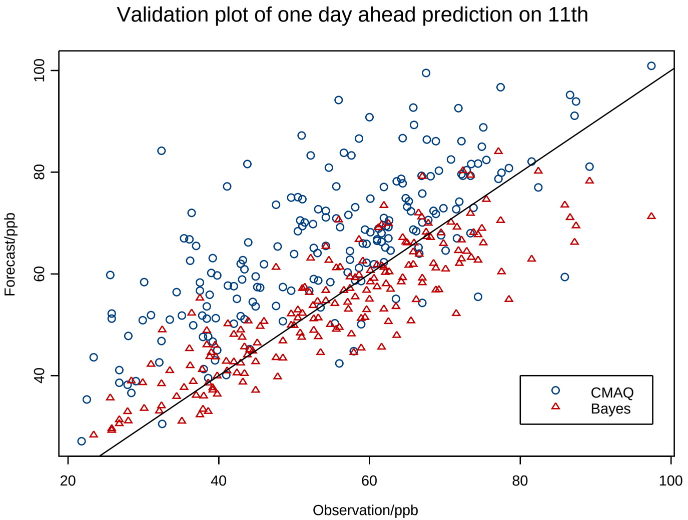
Prediction Quality

Comparison of root mean square error (ppb) (MSE) and relative bias (ppb) (rBIAS)

	RMSE		rBIAS	
Validation Days	CMAQ	Bayes	CMAQ	Bayes
Aug 2–9	15.15	7.47	0.1588	-0.0042
Aug 3–10	15.70	7.20	0.1687	-0.0070
Aug 4–11	16.14	8.03	0.1732	-0.0174
Aug 5–12	15.92	7.51	0.1728	-0.0215
Aug 6–13	15.51	6.53	0.1724	-0.0083

Validation Plot

Validation plot for one day ahead forecast on 11th Aug



Hit and error percentages for O_3 exceeding 80 ppb.

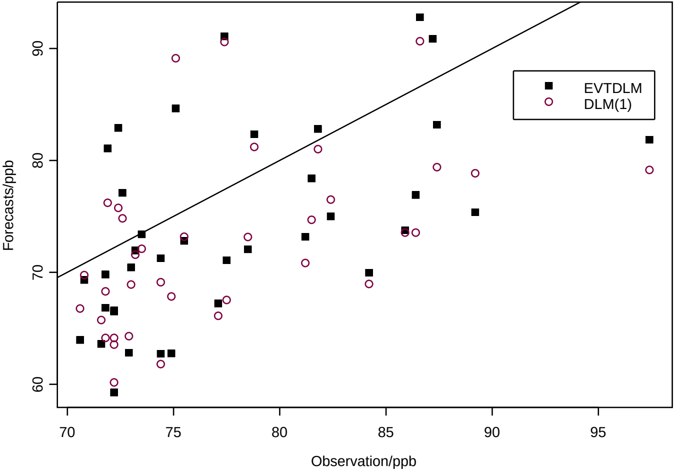
Period	CMAQ Hit	Error	Bayes Hit	Error
Aug 2-9	84.76	15.24	95.12	4.88
Aug 3-10	82.20	17.80	94.24	5.76
Aug 4-11	82.05	17.95	94.36	5.64
Aug 5-12	84.78	15.22	94.92	5.08
Aug 6-13	83.92	16.08	93.97	6.03

Extreme Value Theory Extention

- Not accurate to predict high values (> 80ppb).
- Non-normal distribution.

1. Measurement Equation: $Z(\mathbf{s}, t) \sim GEV(\mu(\mathbf{s}, t), \sigma_g, \nu)$.
2. Second Equation: $\mu(\mathbf{s}, t) = O(\mathbf{s}, t) + \epsilon(\mathbf{s}, t)$.
3. System Equation:
   $O(\mathbf{s}_i, t) = \xi_t + \rho O(\mathbf{s}_i, t-1) + \beta_0 x(\mathbf{s}_i, t) + \eta(\mathbf{s}_i, t)$.

Validation of the upper tail on Aug 13th



RMSE of the upper tail on Aug 13th.

Observed Value	DLM(1)	EVTDLM
All	6.94	7.37
> 50	7.26	7.30
> 60	7.64	7.61
> 70	8.59	8.28
> 80	10.53	9.45

Conclusion

- The forecast is consistent, more accurate, faster than running another computer model.
- Maps of probability statement could be produced.
- The approach is general. We also forecast hourly data under the same framework.
- C language code is developed and a simplifed version S-plus package for a faster hourly model has been developed.
- Future work will focus on using monitoring data from different data sources.

Future Work

- Modelling the whole USA is also needed.
- Using other non-normal distributions.
- Other types of spatial correlation structure could be used.
- The speed of forecast could be further improved which is a trade-off between accuracy and time.

Acknowledgements

- EPSRC Doctoral Training Account in University of Southampton.
- Data provided by Dave Holland in USEPA.

References

Fuentes and Raftery (2005) Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. Biometrics, 61 (1).

Zimmerman and Holland (2005) Complementary co-kriging: spatial prediction using data combined from several environmental monitoring networks. Environmetrics, 16.

Jun and Stein (2004) Statistical comparison of observed and CMAQ modeled daily sulfate levels. Atmospheric Environment, 38.

Sahu, Yip and Holland (2009) Improved space-time prediction of daily ozone concentration levels in the eastern U.S. Atmospheric Environment, 43.

Sahu, Yip and Holland (2008) A fast Bayesian method for updating and forecasting hourly ozone levels. Univesity of Southampton , Technical Report.

Harrison and West (1997) Bayesian Forecasting and Dynamic Models. Springer.