Improved Forecasting of Next Day Ozone Concentrations in the Eastern U.S.

David Holland, U.S. EPA
Joint work with S. Sahu and C. Yip (U. of Southhampton)

ASA Joint Statistical Meetings, 2010
Vancouver, Canada

Modeling Goals

- For U.S. EPA AIRNow real-time operational system, provide a forecast spatial map of next day daily 8-hr maximum O_3 concentrations
  - Using gridded numerical model output and point air monitoring data
- Map needed soon after last hourly O_3 measurement of previous day (i.e. 11pm)
- What we have done:
  - Using a small test data set (Aug 2-14, 2005), have developed several forecast maps and validation statistics
  - Evaluated several models based on comparative fits
  - Results appear to be encouraging

Modeling Overview

- Bayesian space-time model uses point monitoring data and gridded eta-CMAQ forecast data in a regression structure avoiding:
  - avoids change of support problem of some fusion approaches; does not require integration of the observed point level monitoring process to the grid level one
  - avoids modeling ALL eta-CMAQ data and overwhelming information in monitoring data;
- Solve ‘change of support’ problem by inferring on spatial process at a point level given information at both the areal and point level

Ozone (O_3) Pollution

- Ground level O_3: adverse health effects, can cause respiratory problems
- O_3 is a secondary pollutant
- Sunlight + VOC + $NO_x \rightarrow O_3$
- Interested in daily 8-hr maximum O_3 levels:
  - Maximum of averages formed by 8 successive hourly O_3 levels in a day

Sources of Spatial Data

O_3 Monitoring Data:

- Hourly data from 390 real-time monitoring sites in the eastern U.S.
  - Calculate maximum of all 8-hr avg's within a day
- Use data from 350 sites for modeling (estimation and forecasting)
- Set aside data from 40 sites for validation
- Test data for 15 days, August 2-14, 2005
- Some missing data

Eta CMAQ Hourly O_3 Forecast Output:

- High resolution 12 km gridded output over the eastern U.S. (~10,000 grid cells)
  - Calculate daily 8-hr maximum
- Includes next day forecast

Use of Data in Model

- Model the daily 8-hr maximum data for a running window of 7 consecutive days during the two weeks

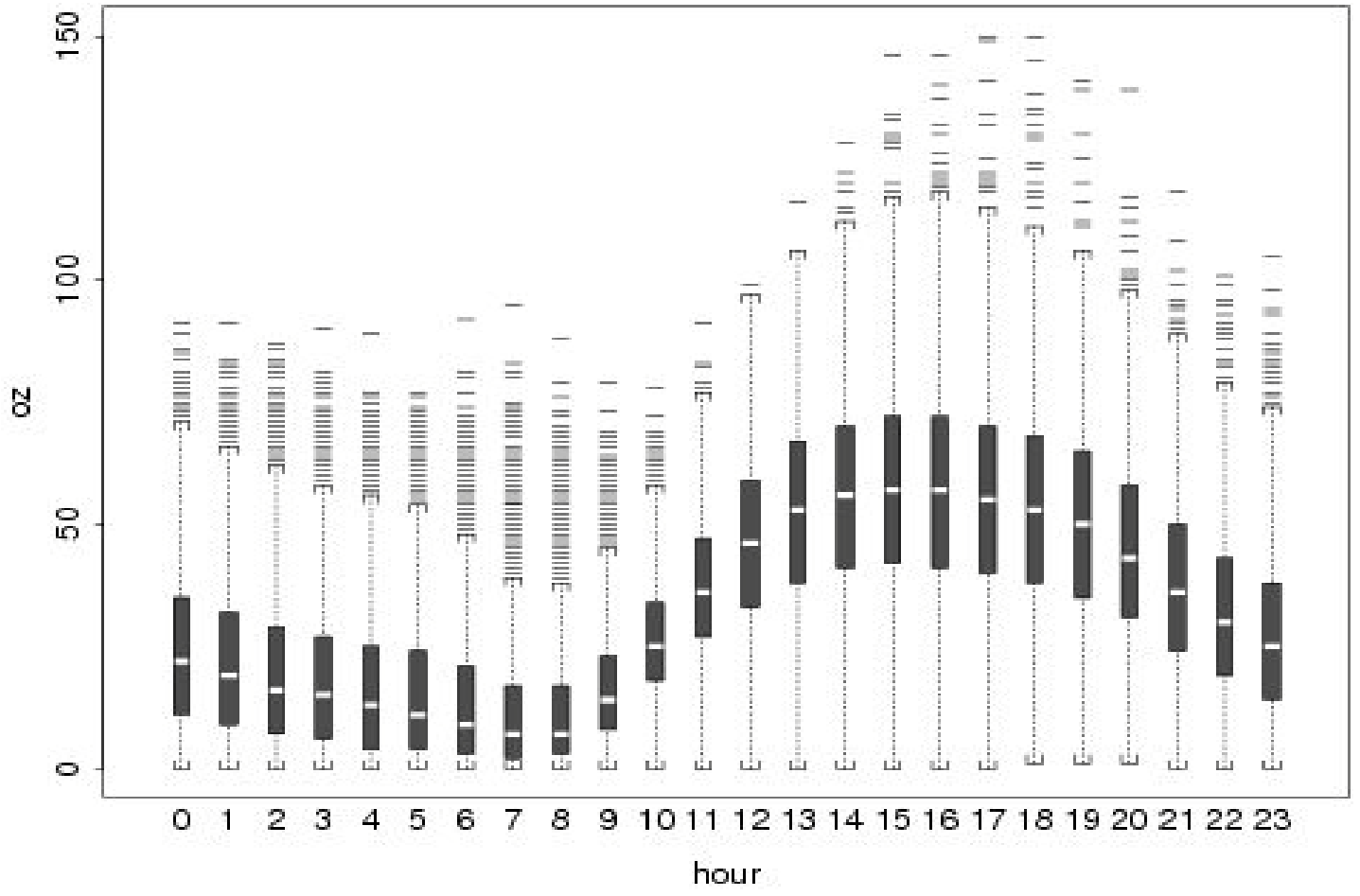
  - weekly cycle arbitrarily chosen, could include more distant data, but initial modeling results show no improvement in forecasts

- Forecast next days spatial pattern of maximum 8-hr average concentrations

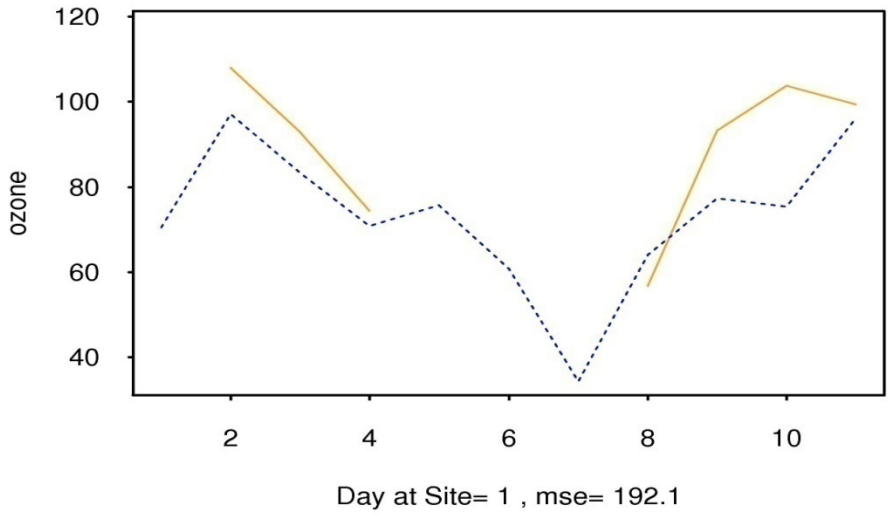
- For each monitoring site, we find the nearest eta CMAQ grid cell centroid to use as a covariate

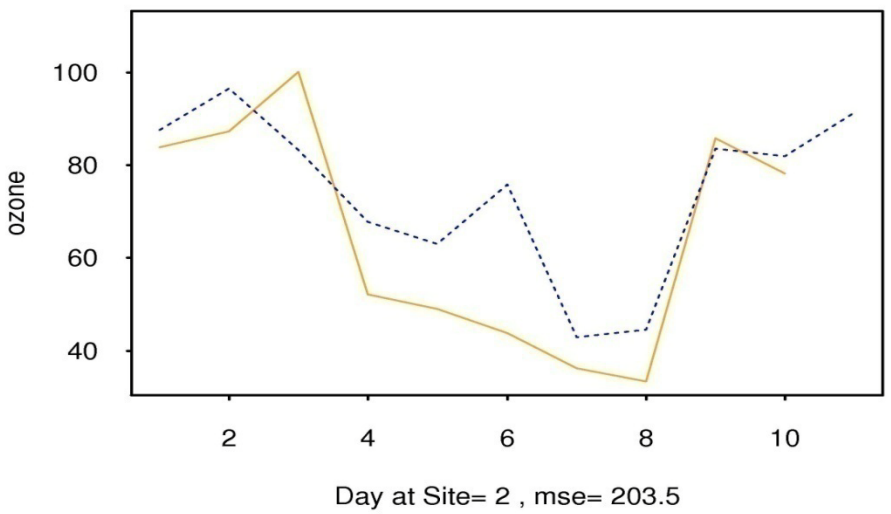
- Finally, randomly sample 3000 CMAQ grid locations to use as a predictive grid; just illustrative for now

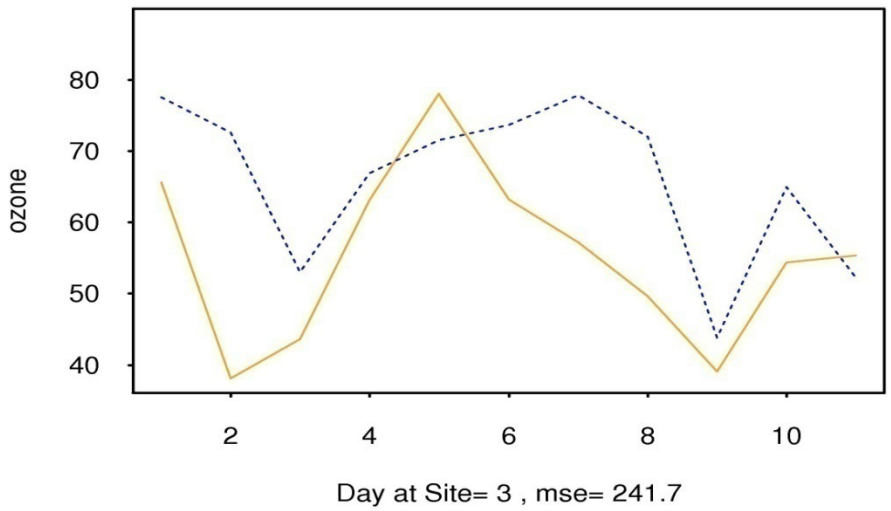
Diurnal Pattern

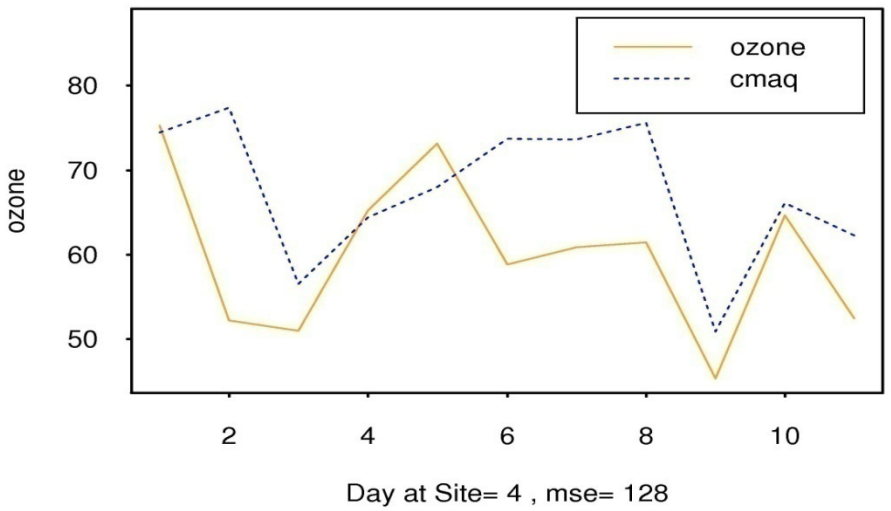


Observed O_3 (tan) vs. Eta-CMAQ O_3 (blue)

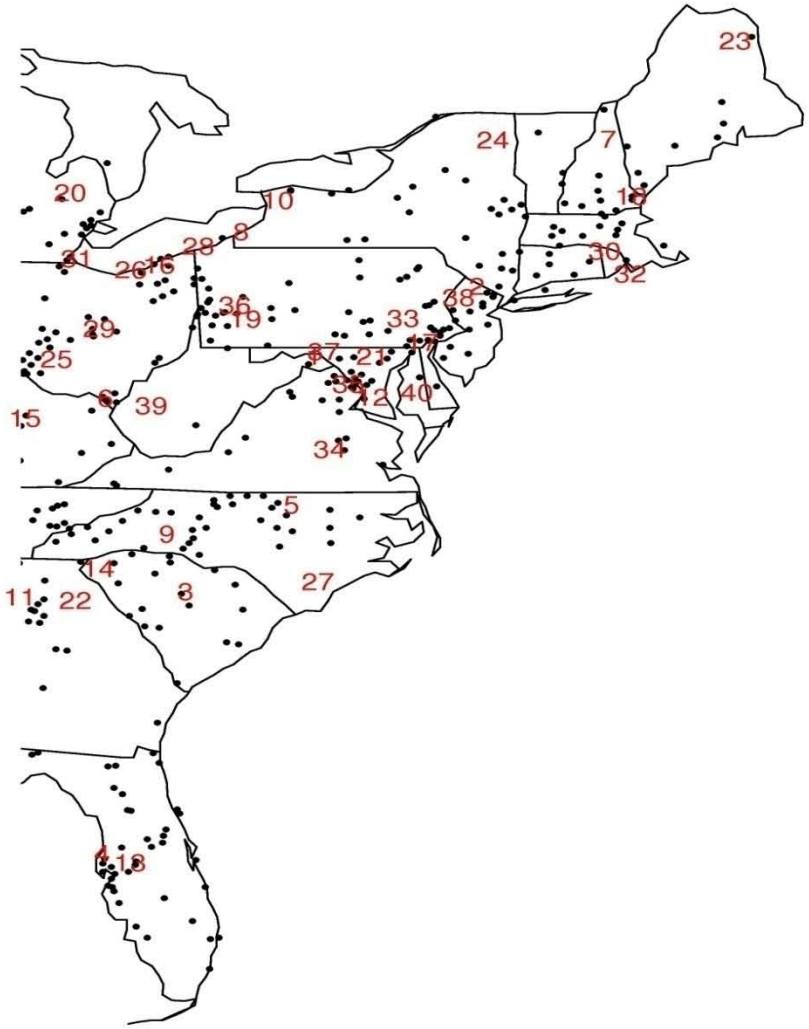


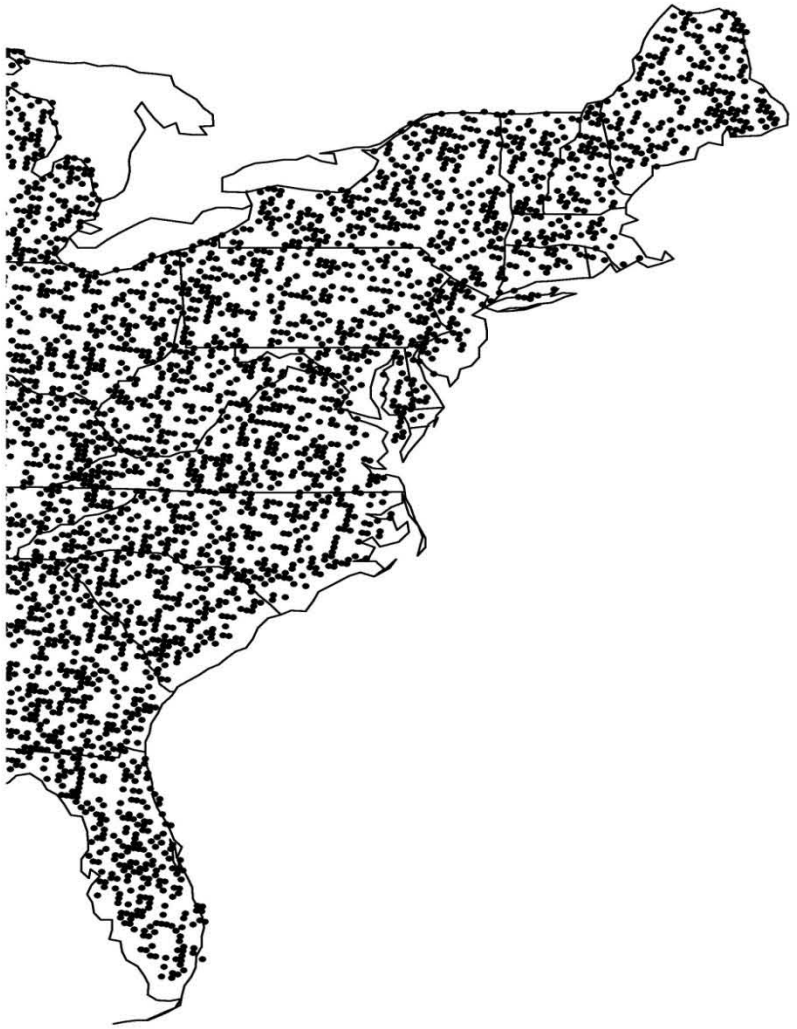






Locations of Model Fitting and Validation sites (left), Predictive grid (right)





Modeling Details

- Model developed here is motivated by Sahu, Gelfand, and Holland (2007), J. American Stat. Assoc., 107, 1221-34.

Square-Root O_3:

- Observed = $Z(\mathbf{s},t)$ at location **s** and time *t*.
- Develop model for n sites denoted by $\mathbf{s}_1,\ldots,\mathbf{s}_n$ for a running window of *t*=1,…,T=7 days
- True square-root O_3 = $O(\mathbf{s},t)$.

Measurement Error Model: Data represent true concentration plus random measurement error

$$Z(\mathbf{s},t) = O(\mathbf{s},t) + \varepsilon(\mathbf{s},t)$$

$$\varepsilon(\mathbf{s},t) \sim N(0,\sigma_\varepsilon^2).$$

for $i = 1,\ldots n;\ \ t = 1,\ldots,T$

$\varepsilon_l(\mathbf{s},t)$ is a white noise process capturing the 'nugget' effect with variance σ_ε^2.

Forecast Model

$$O(\mathbf{s}_i,t) = \xi + \rho O(\mathbf{s}_i,t-1) + \left(\beta_0 + \beta(\mathbf{s}_i^*)\right) x(\mathbf{s}_i^*,t) + \eta(\mathbf{s}_i,t)$$

- ξ is a constant mean across space and time
- $\rho O(\mathbf{s}_i, t\text{-}1)$ is autoregressive ($0 < \rho < 1$)
- $\left(\beta_0 + \beta(\mathbf{s}_i^*)\right)\mathrm{x}(\mathbf{s}_i^*,t)$ spatially varying regression term using Eta-CMAQ as a covariate; $\mathbf{s}_i^*$ denotes CMAQ grid cell containing $\mathbf{s_i}$
  - $\beta \sim N(\mathbf{0}, \Sigma_\beta)$, where $\Sigma_\beta(i,j) = \sigma_\beta^2 \rho_\beta(\mathbf{s}_i - \mathbf{s}_j; \phi_\beta)$
- $\eta(\mathbf{s}_i,t)$ is a spatially correlated, but independent in time error term.
  - $\eta \sim N(\mathbf{0}, \Sigma_\eta)$, where $\Sigma_\eta(i,j) = \sigma_\eta^2 \rho_\eta(\mathbf{s}_i - \mathbf{s}_j; \phi_\eta)$
- Use cross-validation method to decide on including $\beta(\mathbf{s}_i^*)$

Inference Using the Posterior

Using vector notation, e.g. $\mathbf{O}_t = (O(\mathbf{s}_1,t),...,O(\mathbf{s}_n,t))'$

Let $\vartheta_t = \xi\mathbf{1} + \rho\mathbf{O}_{t-1} + \beta_0\mathbf{x}_t + \mathrm{X}_t\boldsymbol{\beta}$ for t=1,…,T, $\boldsymbol{\theta} = \{\beta_0, \boldsymbol{\beta}, \rho, \sigma_\varepsilon^2, \sigma_\eta^2, \sigma_\beta^2, \xi\}$

w denote all unknowns: $\mathbf{O}_t$, missing data $z^*(\mathbf{s}_i,t)$, and non-missing data $z(\mathbf{s}_i,t)$

Then, the log of the posterior distribution, $\log\pi(\boldsymbol{\theta},\mathbf{w}\,|\,\mathbf{z})$, is

$$-\frac{nT}{2}\log(\sigma_\varepsilon^2) - \frac{1}{2\sigma_\varepsilon^2}\sum_{t=1}^{T}(\mathbf{Z}_t - \mathbf{O}_t)'(\mathbf{Z}_t - \mathbf{O}_t) - \frac{nT}{2}\log(\sigma_\beta^2)$$

$$-\frac{1}{2\sigma_\eta^2}\sum_{t=1}^{T}(\mathbf{O}_t - \vartheta_t)'S_\eta^{-1}(\mathbf{O}_t - \vartheta_t) - \frac{nT}{2}\mathrm{lo}\ (\sigma_\beta^2) - \frac{1}{2\sigma_\beta^2}\boldsymbol{\beta}'S_\beta^{-1}\boldsymbol{\beta} + \log(\pi(\boldsymbol{\theta}))$$

ξ, β_0 are independently $N(0,10^4)$, $\rho \sim \mathrm{N}(0,10^4)\mathrm{I}(0<\rho<1)$

$1/\sigma_\varepsilon^2, 1/\sigma_\eta^2, 1/\sigma_\beta^2$ are independently $G(2,1)$ to have a proper prior

Prediction Details

- First predict at any new locations in $t = 1, \ldots, T$
- From measurement error model:

$$Z(\boldsymbol{s}',t) \sim N\left(O(\boldsymbol{s}',\mathrm{t}),\sigma_{\varepsilon}^{2}\right)$$

  - $O(\mathbf{s}',t+1?)$ can only be sequentially determined using previous $O(\mathbf{s}',t)$ up to time t.

- Posterior predictive distribution of $Z(\mathbf{s}',t)$ is obtained by Integrating over all unknown quantities:

$$\pi\left(Z(\mathbf{s}',t)\,|\,\mathbf{z}\right) = \int \pi\left(Z(\mathbf{s}',t)\,|\,O(\mathbf{s}',[t],\sigma_{\varepsilon}^{2}\right)\pi\left(O(\mathbf{s}',[t]\,|\,\beta(\mathbf{s}'),\boldsymbol{\theta},\mathbf{w}\right)$$

$$\times\pi\left(\beta(\mathbf{s}')\,|\,\boldsymbol{\theta}\right)\,dO(\mathbf{s}',[t])\,d\beta(\mathbf{s}')\,d\boldsymbol{\theta}\,d\mathbf{w}$$

Prediction (cont.)

- MCMC methods are used to sample the posterior
- Draws from the posterior distribution $\pi(\boldsymbol{\theta} \mid \mathbf{z}, \mathbf{w})$ and the conditional distribution $\pi(\beta(\mathbf{s}') \mid \boldsymbol{\theta})$ facilitate evaluating the integral in the joint posterior

- In summary, we implement the following algorithm to predict $Z(\mathbf{s}', t), t = 1, ..., T.$

1. Draw a sample $\boldsymbol{\theta}^{(j)}, \mathbf{w}^{(j)}$, $j \geq 1$ from the posterior
2. Draw $\beta^{(j)}(\mathbf{s}')$ using $\beta(\boldsymbol{s}' \mid \boldsymbol{\theta})$
3. Draw $\mathbf{O}^{(j)}(\mathbf{s}', [t])$ sequentially using $\mathbf{O}(\mathbf{s}', t) \mid \beta(\mathbf{s}'), \mathbf{O}_t, \boldsymbol{\theta}, \mathbf{w}$

   - Note that the initial value $O^{(j)}(\mathbf{s}', 0)$ is a constant for all $\mathbf{s}'$

Prediction (cont.)

4. Finally draw $Z^{(j)}(\mathbf{s}',t)$ from $N\left(O^{(j)}(\mathbf{s}',t),\sigma_{\varepsilon}^{2(j)}\right)$

Median is used as a summary measure to preserve 1:1 relationship between O, Z and O^2, Z^2.

Analysis/Results

- Under weak prior assumptions, cannot estimate all covariance parameters, $\sigma_\varepsilon^2, \sigma_\eta^2, \sigma_\beta^2, \phi_\eta, \phi_\beta$, consistently:
  - hence, we use set-aside data to select decay parameters ϕ_η, ϕ_β
  - use a 2-dimensional grid of reasonable ϕ_η, ϕ_β values to optimize MSE
- Only a few $\beta(\mathbf{s}_i)$ were significant:
  - Based on Bayesian model selection criterion of Gelfand and Ghosh (1998), we decided to exclude this term
  - any sort of local lack-of-fit may be compensated by $\eta(\mathbf{s}_i, t)$
- Does not mean that eta-CMAQ has no spatial-temporal bias; if *goal is to estimate bias*, consider eliminating $\rho O(\mathbf{s}_i, t-1)$ from model

Mean-square errors

Validation Days	Eta CMAQ	$\beta(\mathbf{s})\neq 0$	$\beta(\mathbf{s})=0$
Aug 2-9	229.6	84.4	50.5
Aug 3-10	246.4	58	50
Aug 4-11	260.5	77.8	64.5
Aug 5-12	253.4	99.1	62.1
Aug 6-13	240.6	72.5	45.4

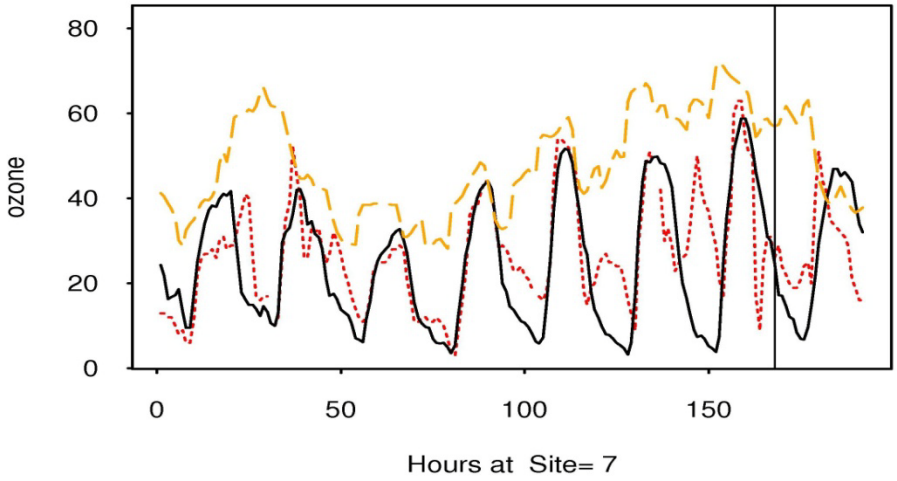
Hit and False Alarm Rates

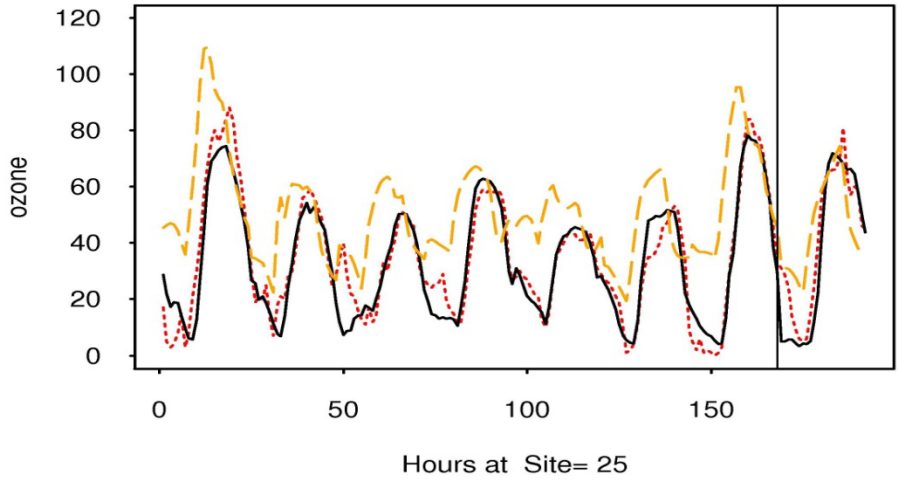
- Hit: defined as event where both the validation observation and forecast are either greater or less than 75 *ppb*
- False Alarm: observation is less than 75 *ppb*, but forecast is greater than 75 *ppb*

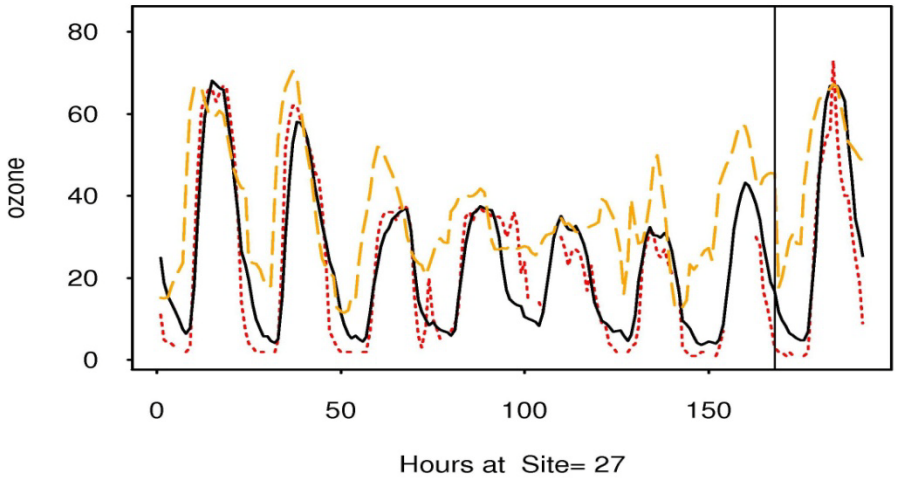
Hit and False Alarm percentages for O_3 exceeding 75 *ppb*

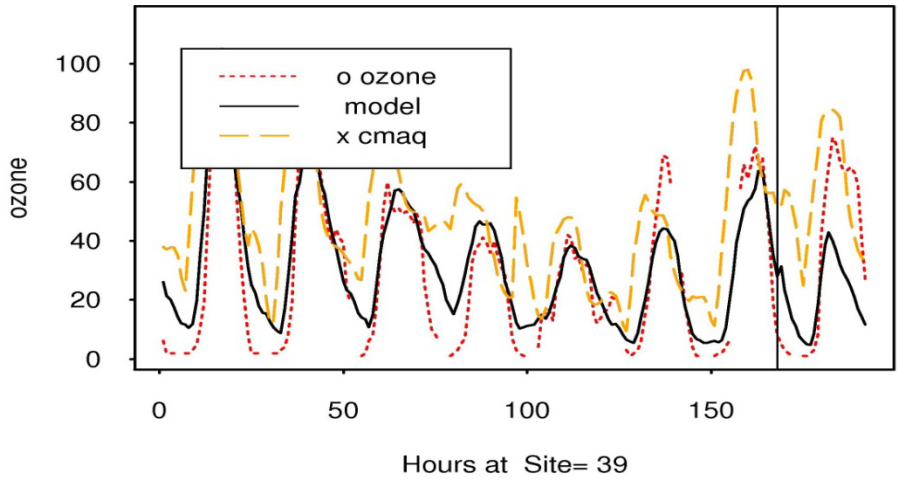
	Eta-CMAQ		Model: $\beta(\mathbf{s}) = 0$	
	Hit	False Alarm	Hit	False Alarm
Aug 2-9	81.7	17.1	90.9	4.9
Aug 3-10	79.6	19.4	92.7	3.7
Aug 4-11	78.9	20.0	93.9	2.6
Aug 5-12	80.7	18.8	93.4	1.5
Aug 6-13	79.9	19.6	93.0	2.5

Hourly Validation, Aug 11
obs data (red), Eta-CMAQ (tan), Forecast (black)

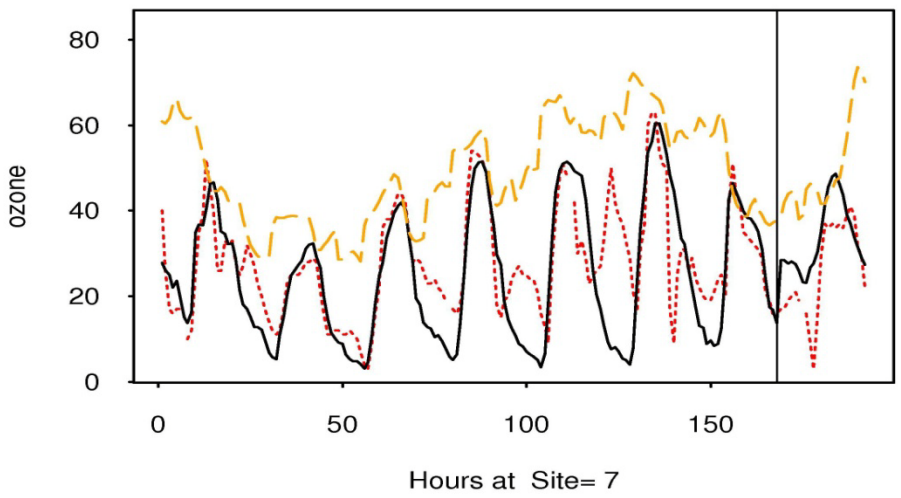


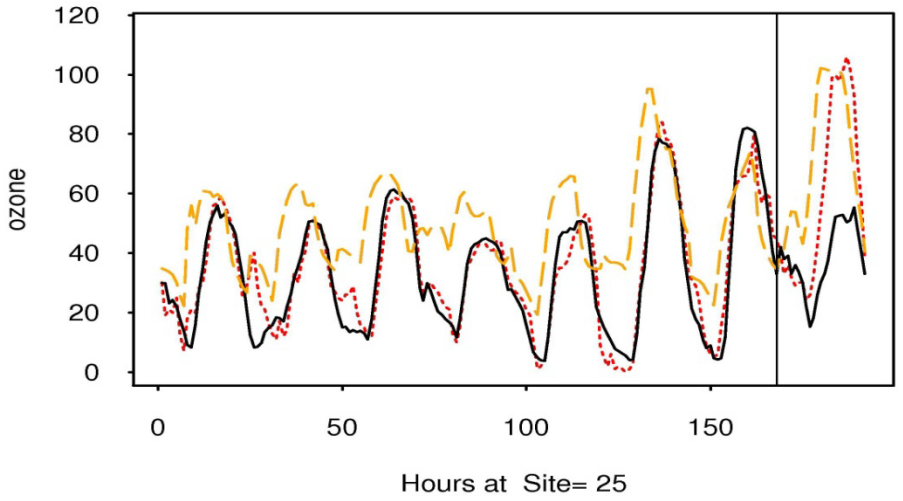


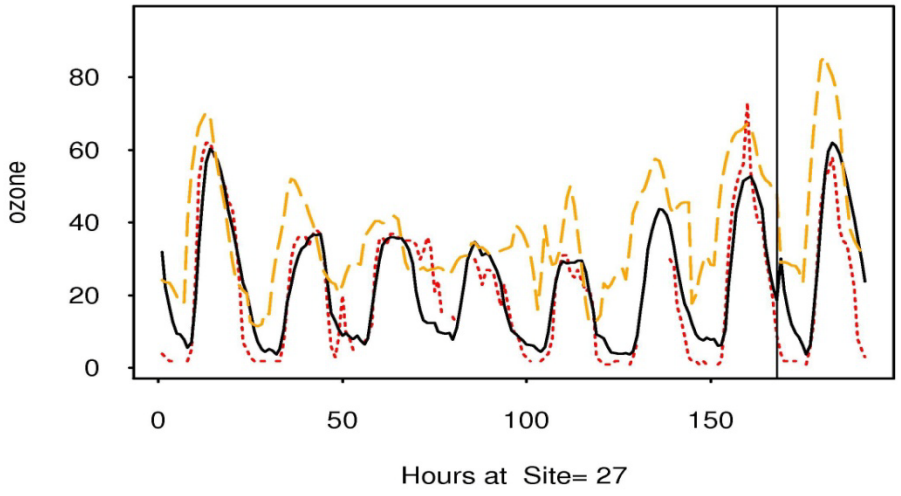


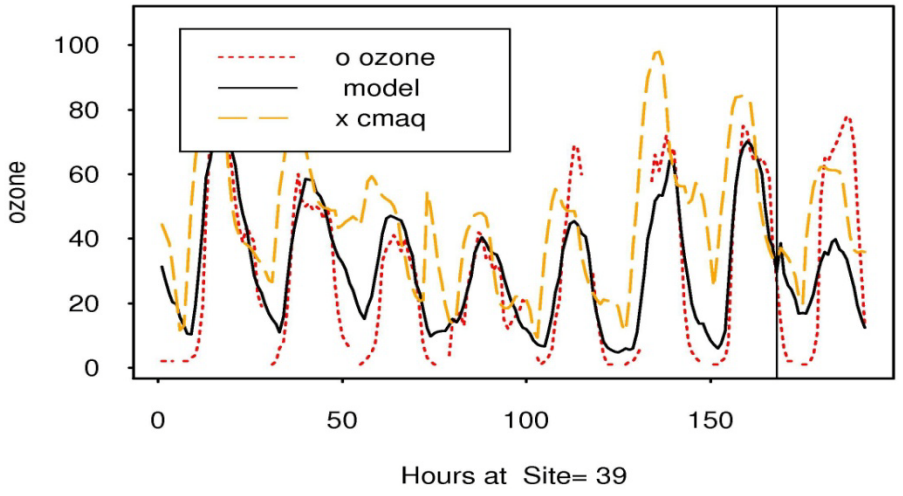


Hourly Validation, Aug 11
obs data (red), Eta-CMAQ (tan), Forecast (black)

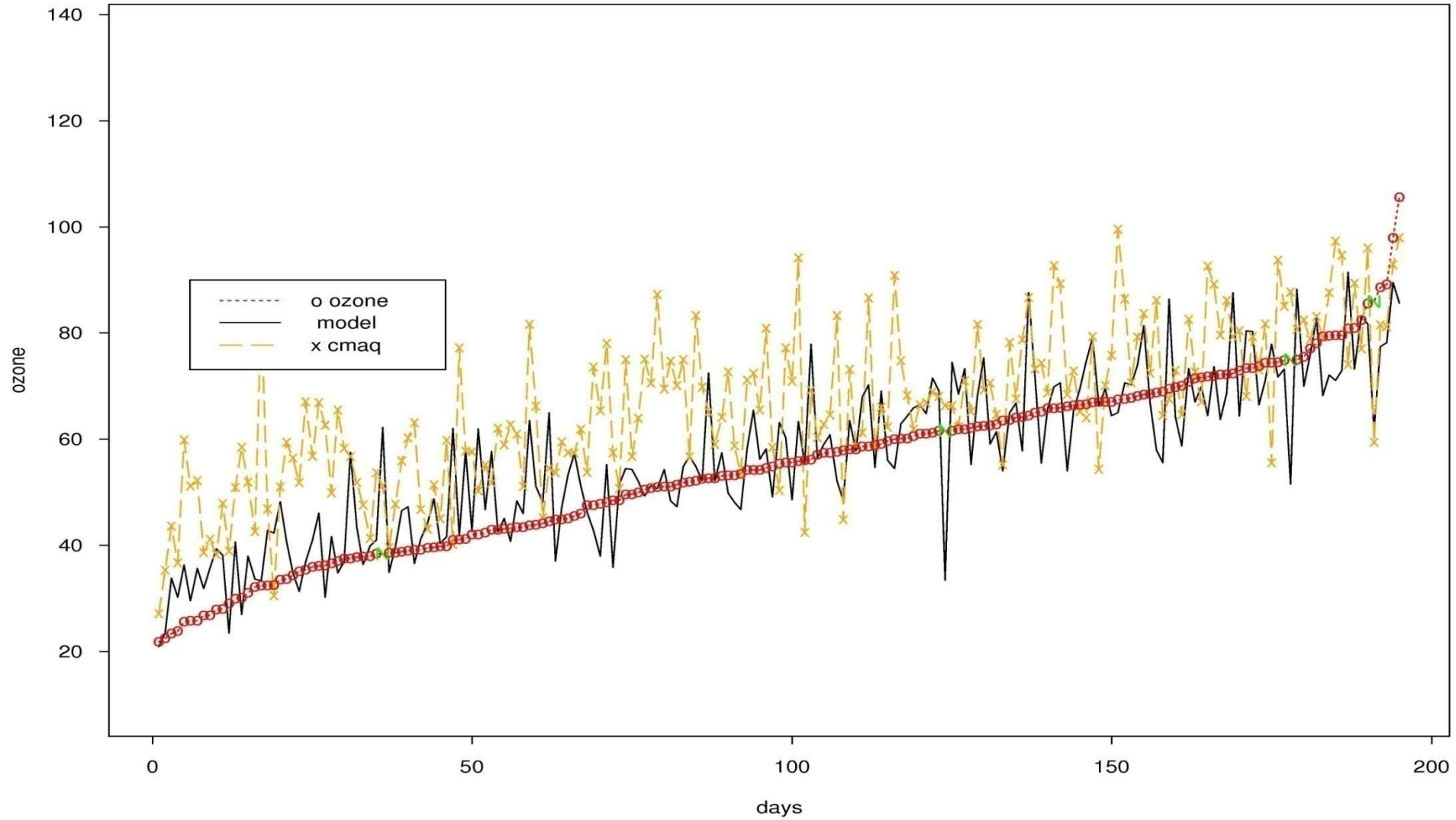




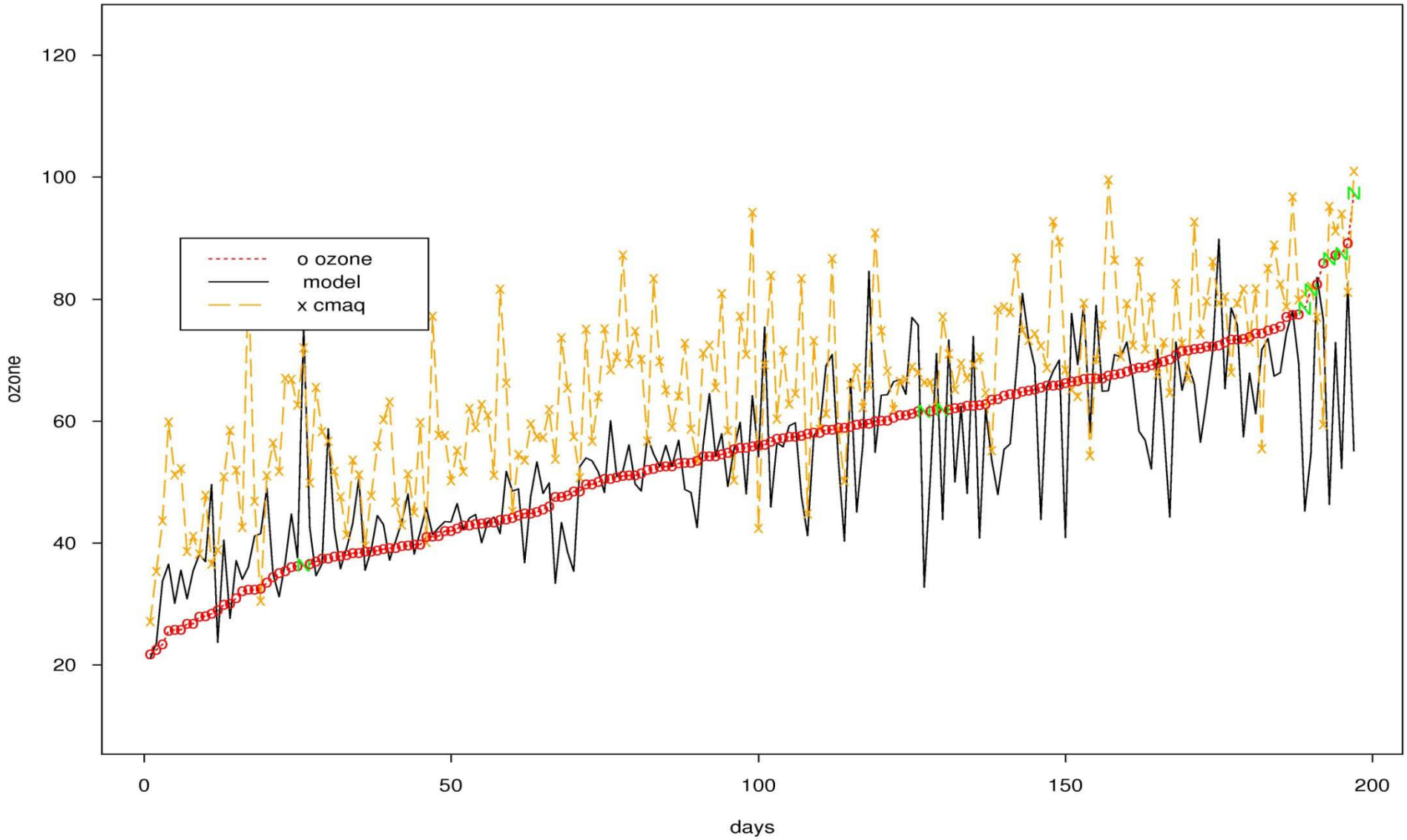




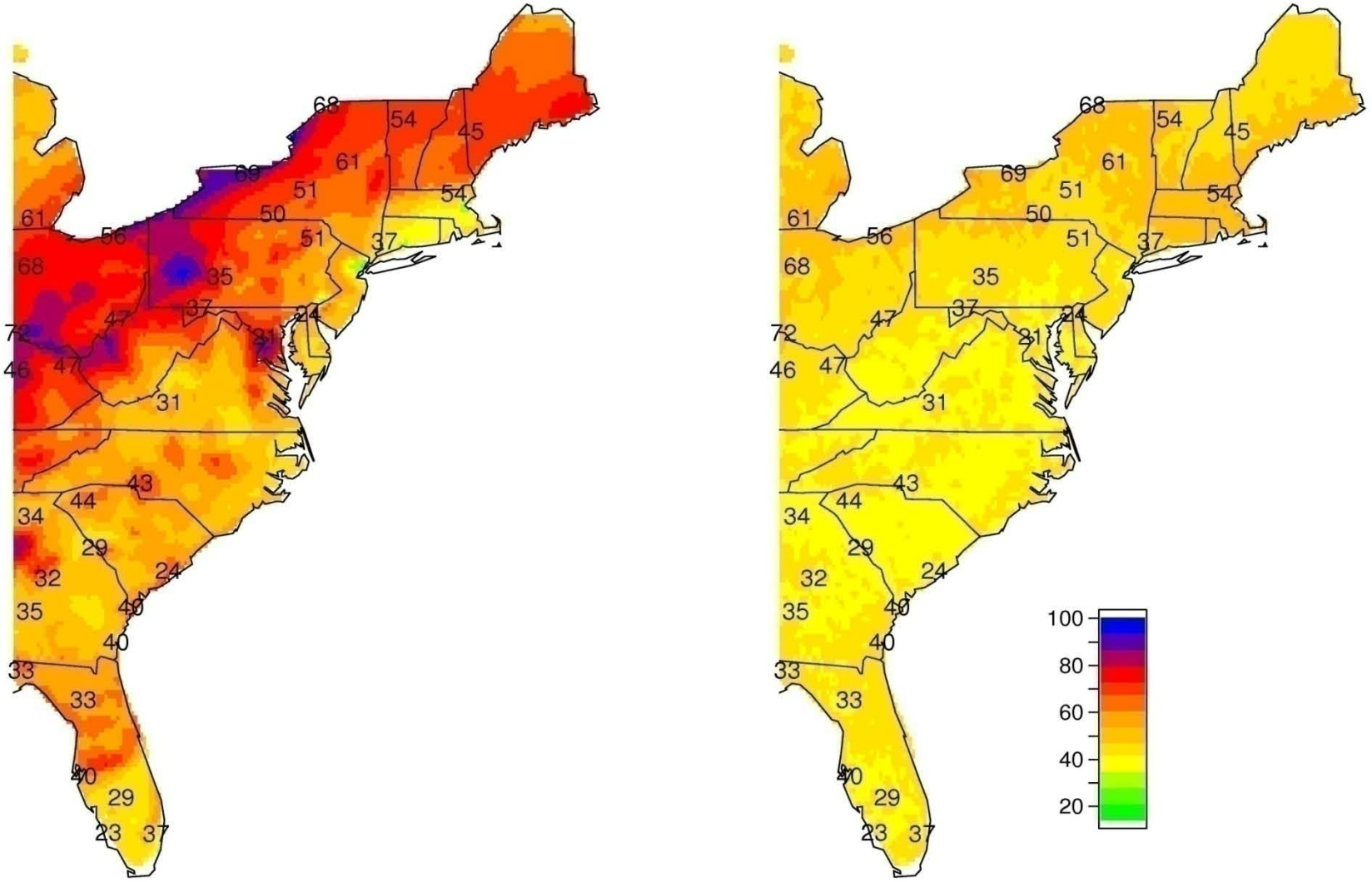
Validation: daily to Aug 11
data (red), Eta-CMAQ (tan), Forecast (black)



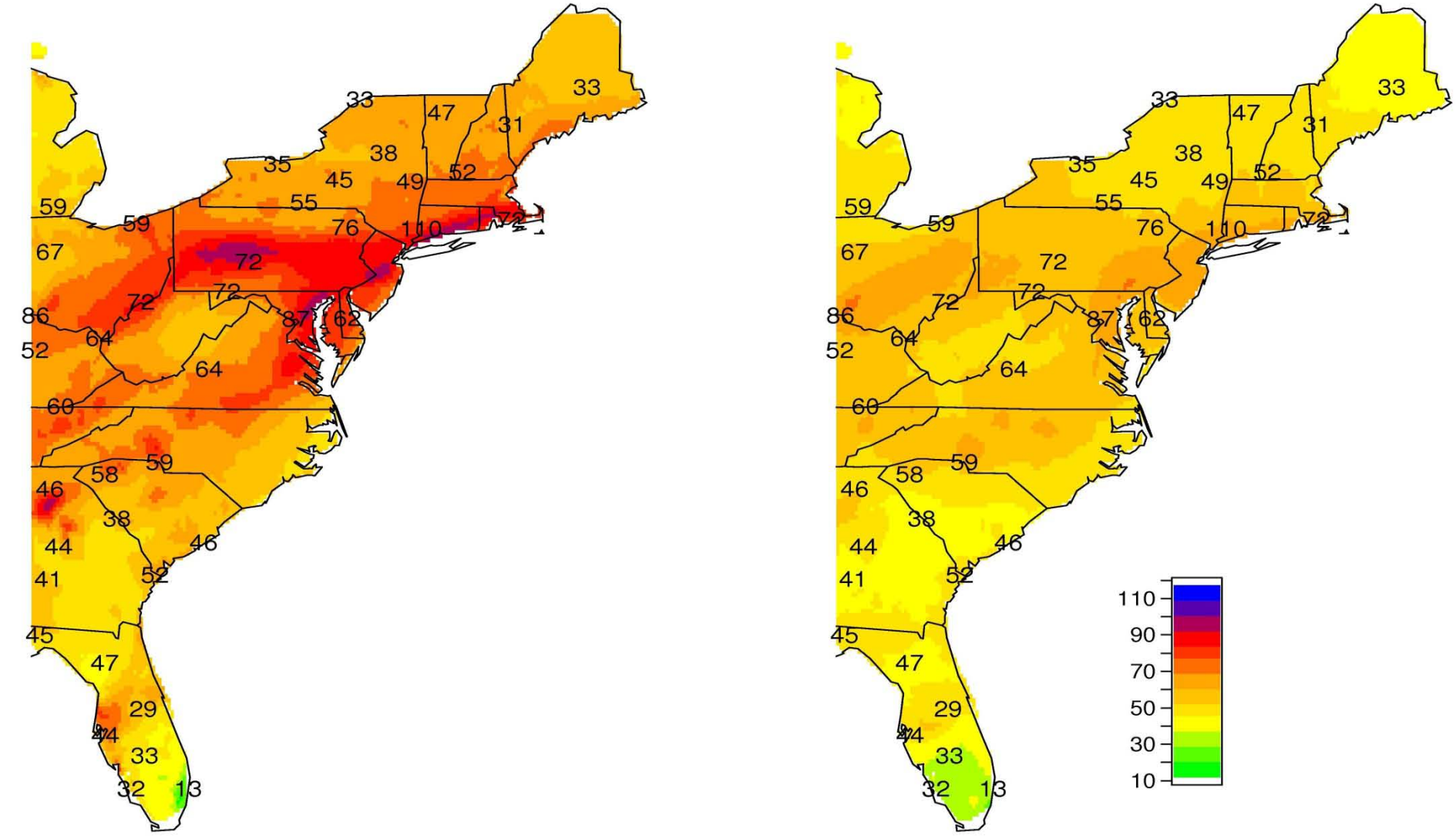
Validation: daily to Aug 12
data (red), Eta-CMAQ (tan), Forecast (black)



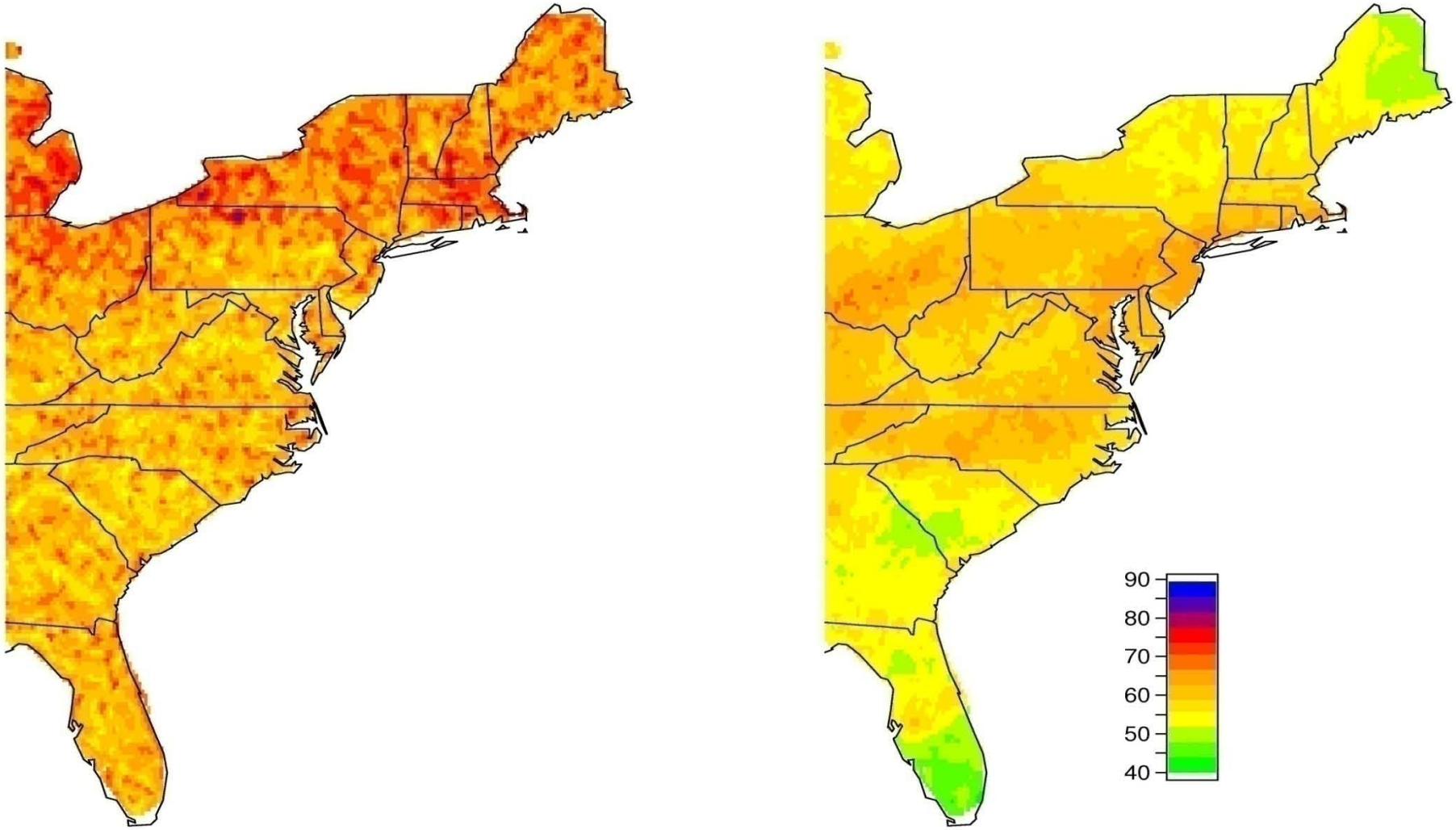
Forecast Maps for August 9
Eta-CMAQ (left), Forecast model $\beta(s)=0$ (right)



Forecast Maps for August 9
Eta-CMAQ (left), Forecast model $\beta(s)=0$ (right)



Lengths of 95% Confidence intervals for forecasts on Aug 9 (left) and Aug 12 (right)



Conclusions/Future Directions

- Model improves upon forecast results from Eta-CMAQ output
- High resolution model seems adequate for fast EPA AirNow implemention
- Can attach prediction uncertainties to forecasts
- In process of developing a real-time forecast system using this model
- Companion effort, predict current 8-hr average O3 levels based on predictions for each of the 4 previous hours, current hour, and for the next 3 hours