Impact of Spatial Information on Kalman Filtering Prediction

Amanda Muyskens, Chris DeFiglia, Jingtian Bai

September 2, 2018

1 Introduction

Air pollution has become a severe problem across the world. Among the most important pollutants, $PM_{2.5}$ is a kind of pollutant that measures quantities of fine particles with diameter 2.5 mircons or smaller in the air. Human activities such as combustion of fossil fuels from power plants, vehicles and certain industries, as well as natural activities such as forest fire and sandstorms can increase $PM_{2.5}$ in the atmosphere. It contains a mixture of various kinds of hazardous substances in various sizes and shapes. Some of them are emitted directly, while others are generated by the reaction of chemicals emitted from such activities. The fine particles may come from local or regional sources, and may also be carried from other places by wind. Due to its relatively small size, $PM_{2.5}$ is inhalable, and when inhaled, these fine particles can get into lungs and may be absorbed into bloodstream, which can affect both lungs and heart. Higher rates of $PM_{2.5}$ have been linked to acute negative health outcomes such as asthma and bronchitis as well as long-term effects such as heart disease, cancer and even premature death. It has also been linked to negative environmental effects such as low visibility and acid rain. Nowadays, the problem of $PM_{2.5}$ has become even more serious in many developing countries such as China and India.

Because of these negative effects, the United States Environmental Protection Agency (EPA) strictly regulates the creation of $PM_{2.5}$ in many aspects such as developing Particulate Matter Air Quality Standards, and it is of high importance to accurately predict $PM_{2.5}$ in order to help people reduce exposure to it. In pursuit of that goal, the EPA hosts the Community Multiscale Air Quality Modeling System (CMAQ). This numerical air quality model is a development project that combines physical models for emissions, chemistry, and physics in the atmosphere to predict air quality measures, such as $PM_{2.5}$ and ozone. It collects meteorological information and emission rates from sources of emissions that affect air quality, combines meteorological models, emission models, and an air chemistry-transport model to simulate output. So far, EPA has developed CMAQ 5.2 and CMAQ has been widely used by many departments in a various of fields. In addition to the CMAQ model output, EPA's database called AQS receives measurements of pollutant concentrations from over 4000 air quality monitoring stations across the United States where daily values of $PM_{2.5}$ are collected.

However, there are still problems with CMAQ and observational data. On the one hand, CMAQ model is often biased and often has spatial errors when compared to observational data collected from monitoring stations. On the other hand, the spatial coverage of monitoring stations are still quite poor, and many regions cannot find representative monitoring stations.

The purpose of our project is to explore approaches to assimilate the output from CMAQ model with observational data from monitoring stations in order to make accurate predictions on the $PM_{2.5}$ concentration. In particular, we aim to explore if performances will be improved by employing Kalman Filter. This problem is particularly challenging because the CMAQ model output (in grid cell) and the observational data are not based on the same spatial locations, and it is not clear how to define the covariance matrices necessary to apply the Kalman Filter (P_0^f, Q_k, R_k).

There are many researches focusing on assimilating the CMAQ model output with observational data. For example, Choi et al. (2009) combine daily average $PM_{2.5}$ concentration from observational data and computer model output to predict true daily average concentration aggregated over counties at each time point. McMillan et al. (2009) combine daily average $PM_{2.5}$ concentration from monitoring stations and CMAQ model output to predict true daily average

concentration for each day in 2001. Berrocal et al. (2010) combine daily 8-hr max ozone concentration from observational data and CMAQ model output to downscale the model output at point level. Actually, there is much debate in the air quality literature how the CMAQ model output should be incorporated into prediction, and we propose here use of the Kalman Filter for data assimilation, which is different from most previous researches. Since the model itself should be incorporated when implementing the Kalman Filter, we perform simulation studies by making assumptions about the true model instead of performing the real data analysis, which will be explained in the next section. In addition, because of the high dimensional aspect of CMAQ, many use only the closest observation or a window of observations around the monitoring stations for prediction, which will also be discussed in more details in later sections.

2 Simulation Methods

We consider the above motivation and consider applying the Kalman filter to data simulated as the motivating air quality data and perform various simulations to explore the impact of treatment of spatial correlation on prediction. In both CMAQ and monitoring data, we simulate log $PM_{2.5}$ estimates, but at differing spatial locations. CMAQ output is the combination of complex mathematical models and therefore we do not have an explicit differential equation to use that describes the dynamic movement of the model. Then, instead, we adopt the space-time separable statistical model, where time is an autoregressive(1) process and space follows an exponential covariance.

Let X_k be the $n \times 1$ vector of model data at all spatial locations at time k and Y_k be the $m \times 1$ vector of model data at all monitoring site spatial locations at time k. Note that $n >> m$. We assume all points in X and Y follow a Gaussian process and therefore all the points can be written as jointly multivariate normal. We define their unconditional distributions as

$$\begin{pmatrix} X_0 \\ X_1 \\ \vdots \\ X_k \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_x & \rho\Sigma_x & \dots & \rho^k\Sigma_x \\ \rho\Sigma_x & \Sigma_x & \dots & \rho^{k-1}\Sigma_x \\ \vdots & \vdots & \ddots & \vdots \\ \rho^k\Sigma_x & \rho^{k-1}\Sigma_x & \dots & \Sigma_x \end{pmatrix}\right)$$

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_k \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_y & \rho\Sigma_y & \dots & \rho^{k-1}\Sigma_y \\ \rho\Sigma_y & \Sigma_y & \dots & \rho^{k-2}\Sigma_y \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{k-1}\Sigma_y & \rho^{k-2}\Sigma_y & \dots & \Sigma_y \end{pmatrix}\right)$$

$$\begin{pmatrix} X_k \\ Y_k \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_x & \beta\Sigma_{xy} \\ \beta\Sigma_{yx} & \Sigma_y \end{pmatrix}\right)$$

where we assume $\Sigma_x, \Sigma_y, \Sigma_{xy}$ are stationary, isotropic, exponential spatial covariances. That means that the covariance between points at spatial locations s_i and s_j is defined as

$$\Sigma[i,j] = \sigma^2 \exp\left(-\frac{||s_i - s_j||^2}{\phi}\right)$$

With fixed spatial locations, this high dimensional problem is summarized by two covariance parameters σ^2, ϕ. Then, we derive the conditional distributions necessary for the Kalman filter as

$$p(X_{k+1}|X_k) \sim N(\rho X_k, (1-\rho^2)\Sigma_x)$$
$$p(Y_k|X_k) \sim N(\beta\Sigma_{yx}\Sigma_x^{-1}X_k, \Sigma_y - \beta^2\Sigma_{yx}\Sigma_x^{-1}\Sigma_{xy})$$

Unlike many problems we studied in class, the linearity and normality assumptions are not a problem here because these assumptions are also justified in the data generation. The primary basis for parameteric spatial-temporal statistics is the assumption of a Gaussian Process, meaning any two points in the domain are assumed to be jointly normal. Therefore, instead here we hope to evaluate how incorporation of the model data can improve one day ahead prediction of $PM_{2.5}$

Then, assuming the above models, the full spatial Kalman filter is

Analysis:

$$X_k^a = x_k^f + K_k(Y_k - \beta\Sigma_{yx}\Sigma_x^{-1}x_k^f)$$

$$P_k^a = (I - K_k \beta \Sigma_{yx} \Sigma_x^{-1}) P_k^f$$

$$K_k = P_k^f (\beta \Sigma_{yx} \Sigma_x^{-1})^T (\beta \Sigma_{yx} \Sigma_x^{-1} P_k^f (\beta \Sigma_{yx} \Sigma_x^{-1})^T + \Sigma_y - \beta^2 \Sigma_{yx} \Sigma_x^{-1} \Sigma_{xy})^{-1}$$

Forecast:

$$x_{k+1}^f = \rho X_k^a$$

$$P_{k+1}^f = \rho^2 P_k^a + (1 - \rho^2) P_0$$

Because the CMAQ model is high dimensional since it contains fine spatial griding, there is debate in the literature as to how to incorporate the model output for prediction. There are three general treatments of the space and therefore we apply the Kalman filter under these three set of spatial assumptions.

1. The first and most obvious way to combine the model data and the observational data is to use all the data and the model as stated above with full spatial covariance matrices. This gives us good spatial coverage and therefore prediction both at the observed locations and new locations are ideal. However, this is often not utilized because the CMAQ output is very dense and computing is often difficult or impossible. Although we assume a stationary covariance structure in this project, this assumption is difficult over a large spatial area such as the United States. Then, by including model output from distant spatial locations, if the covariance is indeed non-stationary, we may be biasing our results.

2. Since the spatial coverage of the model data is very dense, there is not much distance between the monitoring data sites and their closest gridded observation. Therefore, the most common method of incorporation of CMAQ is to choose the closest CMAQ observation and assume it is at the same spatial location of the data. Then, we treat the observations as independent in space. In estimating $PM_{2.5}$ at locations where we observe data, this becomes time series Kalman filter prediction at each spatial location. If we wish to predict $PM_{2.5}$ at new locations, we must assume that the independently predicted values are spatially correlated. Because there are only a few points, which are spatially sparse, we expect predictions at new locations to be poor with this method.

3. As machine learning becomes more popular, more methods like this are becoming commonplace in the spatial statistics literature. The Haas (1990) Moving Window method is popular as of late and we apply the same principles here to the Kalman filter. The general idea here is that only close spatial points are important to prediction. Therefore, we define a window around each individual monitoring station, and use only those observations for individual assimilation and prediction. When we predict at new locations, however, it is not clear which window to use for prediction. We assign the new points to the closest window, however, because their spatial distribution is poor, there may be only a few supporting points and they may not be that spatially close. This method is attractive because it is fast as well as incorporating spatial information.

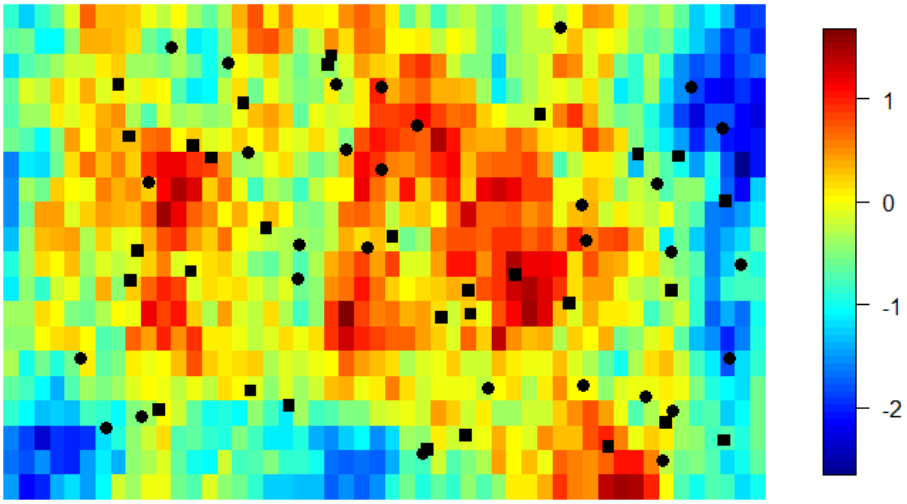
In addition to these three Kalman filter methods, for comparison, we include prediction using only the observed data (Y) as well as a method of only using the model (X).

Ideally, we would like to compare how these methods work in the case where we know the parameters, as well as when they are estimated. However, once we began working on the problem of parameter estimation in this case, it became clear this is a problem fit for further future research. In general, the ensemble Kalman filter could be applicable. However, when we assume a parametric spatial Gaussian process, we hope to draw inference only on a few parameters, rather than the unstructured computation of the sample covariance. Therefore, it would be reasonable to extend the Kalman filter where we use maximum likelihood estimates of the defined parameters in order to offer estimated covariances to impute in the Kalman filter equations. Unfortunately, these maximum likelihood estimates are very difficult to obtain. As we have assumed that these parameters are stationary in time, the correct way to estimate the parameters would be use the unconditional multivariate normal joint distribution of all the data up to that time step and then perform optimization on the parameters. Since there are both X and Y involved, although we have assumed a separable model in each of these variables, the resulting full covariance is not separable. Therefore, MLE estimation would be extremely slow as we would need to calculate the determinant and inverse of a large covariance matrix in each likelihood evaluation. When the time step k becomes even reasonable, this problem becomes computationally impossible. In the

dynamic case where the parameter estimates depend on k, the parameters in the exponential case are shared in both conditional distributions. In addition, even the mean parameters are shared in the covariance and mean specification and then the known GLS estimate of mean parameters is not applicable here. Because of these issues, this is a difficult problem, which one could explore as a dissertation chapter or published paper in the future. Thus, instead we perform sensitivity analysis where we generate data with a set of known parameters and implement our methods with miss-specified parameter estimates to judge the possible impact of estimation.

We design a simulation where we mimic the setting of the air quality data. We generate gridded model data X to mimic CMAQ output on a 20×50 grid. Our objective is to evaluate $PM_{2.5}$ prediction, and we propose two comparisons. First, we consider the criterion of one-day ahead prediction of $PM_{2.5}$ at the locations of the monitoring stations. Second, we consider the one-day ahead prediction of $PM_{2.5}$ to a new set of locations within the spatial domain. Since we do not observe new sites only on locations our assimilated models will exist, this will require statistical kriging using estimated parameters to the new set of locations, given those we do have in our model domain.

We generate this under the aforementioned conditional distribution for a 7 day period, including time 0. Similarly, we generate Y data during this same 7 day period. The monitoring locations are fixed throughout the 7 day period and are sampled as random selection of coordinates with added normal 2-dimensional errors so that they are at different spatial locations than the model grid. We compare the 5 aforementioned prediction methods to estimate observational data for the following day. Below is an image of one example sampled X_0 with the black dots being the monitoring stations where we observe Y. Black squares represent new locations where we would like to predict $PM_{2.5}$.



We explore a variety of scenarios and how these may interact with spatial treatment. We consider different densities of points, different spatial correlations, and a variety of miss-specification of parameters. In specific, we simulate 100 replications of the following scenario modifications. For the baseline, we assume there are 30 monitoring stations, $\rho = 0.8$, $\beta = 0.85$, $\phi = 10$, and $mean = 0$.

1. Truth baseline; No miss-specification
2. Truth $\phi = 5$; No miss-specification
3. Truth $\phi = 20$; No miss-specification
4. Truth number of points = 10; No miss-specification
5. Truth number of points = 90; No miss-specification

6. Truth $mean(Y) = 1$; Miss-specify $mean(Y) = 0$
7. Truth baseline; Miss-specify $\rho = 0.6$
8. Truth baseline; Miss-specify $\rho = 0.95$
9. Truth baseline; Miss-specify $\beta = 0.9$
10. Truth baseline; Miss-specify $\beta = 0.5$
11. Truth baseline; Miss-specify $\phi = 5$
12. Truth baseline; Miss-specify $\phi = 20$

Evaluation of the success of these five methods are evaluated via mean squared prediction error of the observations at the true or new spatial locations. Let Y_{ij} be the true $PM_{2.5}$ observation at spatial location i and on day j, and $\hat{Y_{ij}}$ be the predicted $PM_{2.5}$ for one of our methods. Then,

$$MSE = \sum_{i=1}^{m}\sum_{j=1}^{7}(Y_{ij} - \hat{Y_{ij}})^2$$

. In addition, we compare the computing time to perform each method for one-day ahead prediction over a week. Standard error stated is the simulation standard error. Because we were all more familiar with R, this is the software we use for the project.

3 Results

For this section, the five models will be abbreviated in the plots so as to make them more readable:

- FSKF (Full Spatial Kalman Filter)
- IndKF (Independent Kalman Filter using nearest observation)
- WinKF (Moving Window Kalman Filter)
- KnRho (Predicting with only data, however ρ is known)
- NoDA (No Data Assimilation, simply projecting the model forward)

First let's take a look at the amount of time each model took to run.

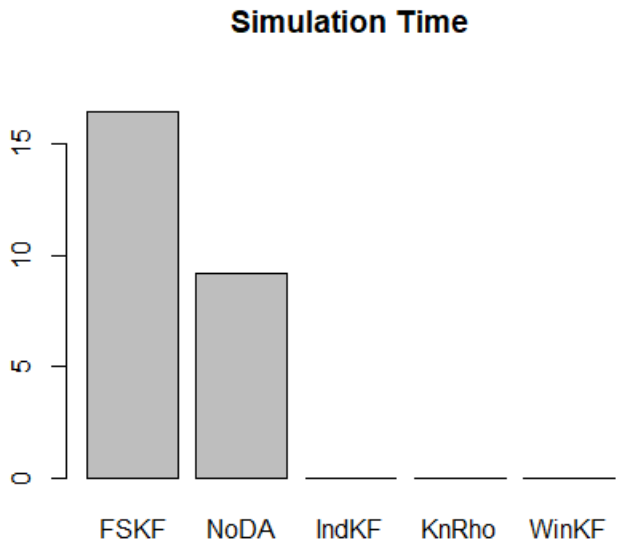


Figure 1: Average simulation time for each of the five models

Both the FSKF and NoDA took a considerable amount of time to run, which isn't too surprising as these methods are working with the entire grid of data all at once. The other three methods take much less time comparatively as they don't attempt this. Even so, it took us about twelve hours total to run the seven day simulation for all twelve settings.

Now, let's look at how each method performed over the seven day period for one simulation:

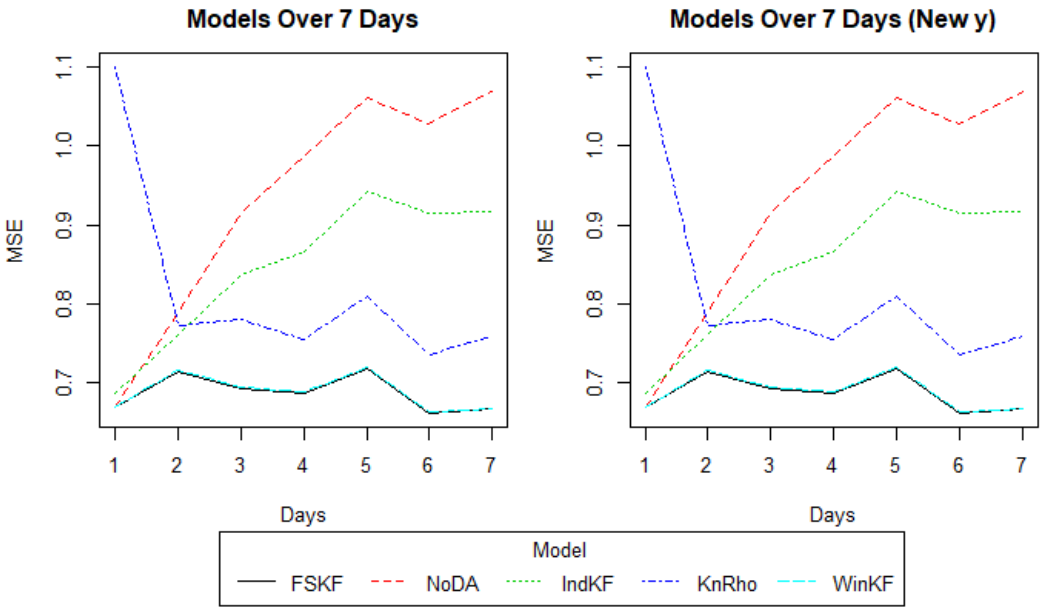


Figure 2: MSE for each model over the course of seven days

In the case without Data Assimilation, our model does worse as time progresses and seems like it will eventually become useless for any kind of accurate prediction. The IndKF model doesn't do amazingly either but the MSE has at least plateaued somewhat by the seventh day. The NoDA model is still increasing by comparison. The KnRho model begins terribly on day 1, but the MSE quickly drops and flattens out. Both the FSKF and WinKF models do the best as they consistently have roughly the same low MSE. WinKF looks to be preferred over all other models as it has a high degree of accuracy and it's fast.

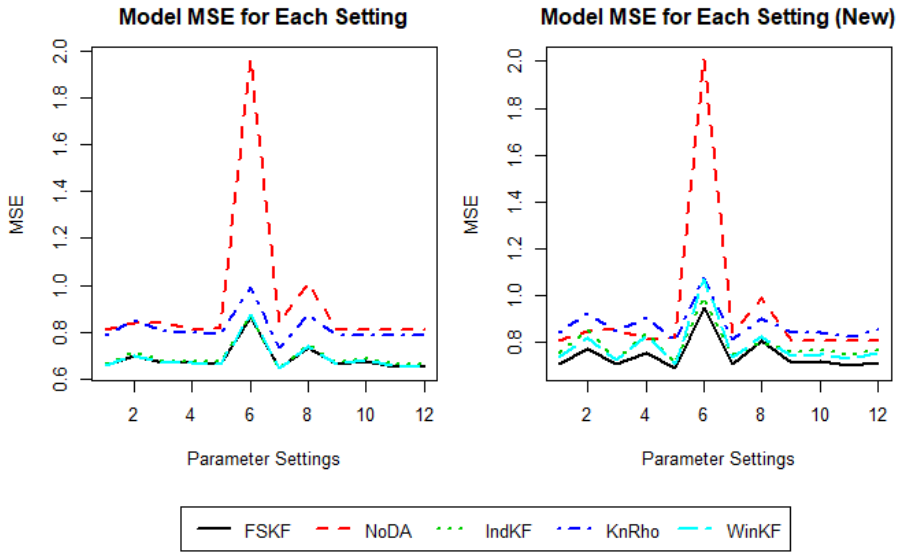


Figure 3: Average model MSE for each of the twelve settings using gridded locations and new locations

Immediately, we see an enormous difference between the NoDA model and all of the others on setting 6, which is where we have a baseline truth $mean(Y) = 1$ and miss-specify $mean(Y) = 0$. With the exception of KnRho on a few settings, all of the Data Assimilation outperform NoDA. On the previous plot, it was somewhat surprising that the average MSE at new locations wasn't noticeably higher than at the grid points, however this doesn't seem to hold true for every one of twelve settings and the MSEs of all models tend to be much more variable at these new points. The FSKF, IndKF, and WinKF have the lowest MSEs overall.

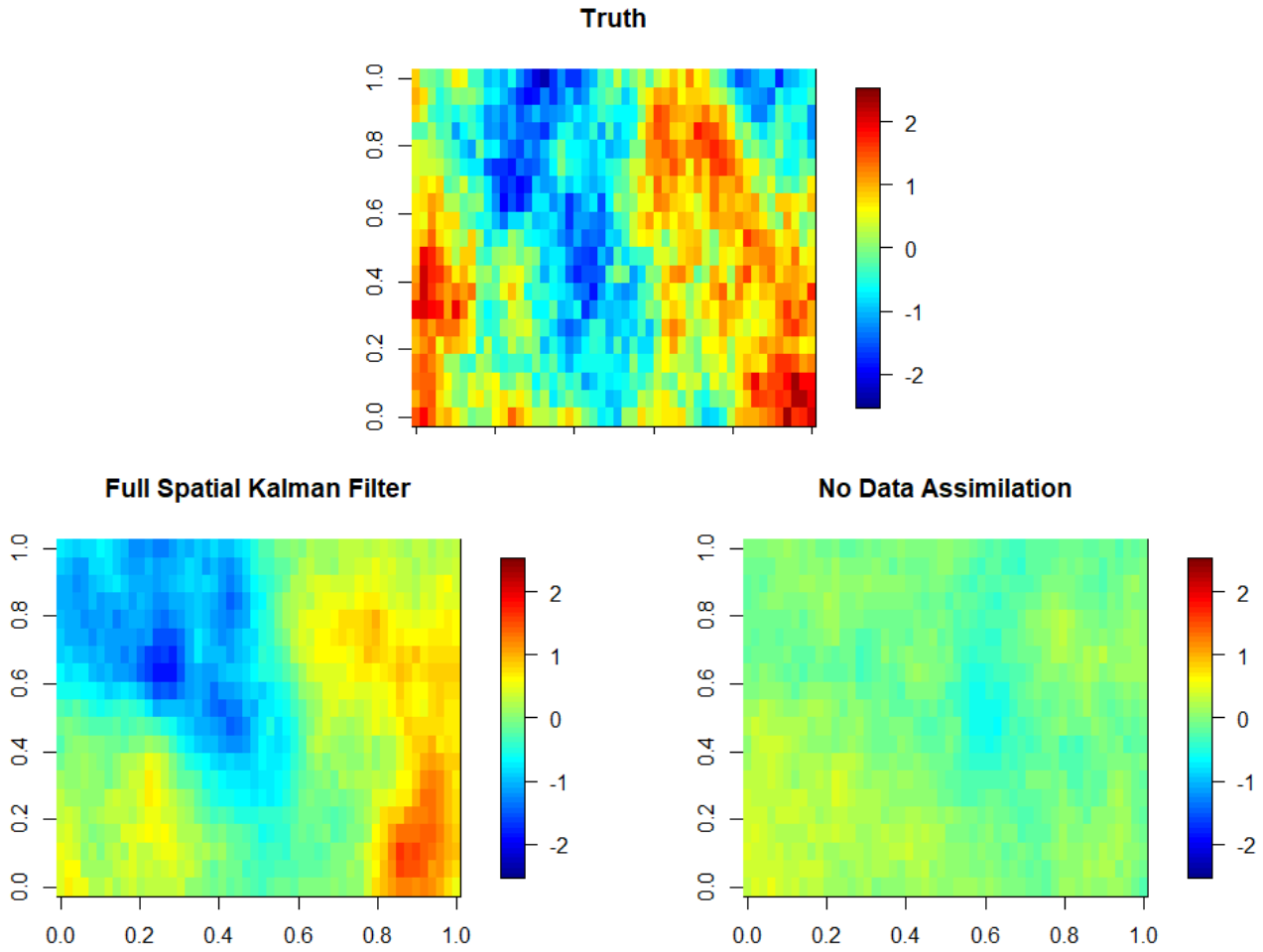


Figure 4: Heat Maps for the Truth, Full Spatial Kalman Filter, and without Data Assimilation on the seventh day.

As opposed to simply looking at MSEs, here we have heat plots which demonstrate the superiority of a model with Data Assimilation over one without it. The NoDA model on the seventh day is almost uniform, without much variation whereas the FSKF succeeds in capturing the major hot and cold regions. The FSKF doesn't completely succeed in predicting the truth as it misses many of the more subtle variations, however it does come fairly close on a macro scale.

4 Conclusions

In this project, we used five models to assimilate simulated CMAQ output and observational data with the purpose of accurately predicting $PM_{2.5}$ concentrations at grid points and new locations. We indeed viewed a significant performance increase by using models that incorporate Data Assimilation over ones that don't over a seven day period. The model we used that simply projected the model forward without DA not only took a considerable amount of time to run, but it also had high error rates. Additionally, we performed a sensitivity analysis on said models as parameter estimation was outside the scope of this project.Out of all the DA models we used, the Moving Window Kalman Filter seemed to be the best given that it runs quickly and it has low error. If computational resources were not an issue, the Full Spatial Kalman Filter would also be an option. Future work could involve applying these models to real data and evaluating their performance as well as estimating the model parameters.

A Simulation Results

Table 1: Simulation Results at Grid Points and New Locations

	1	2	3	4	5	6	7	8	9	10	11	12
FSKF MSE	0.66	0.70	0.67	0.67	0.66	0.86	0.64	0.74	0.66	0.68	0.65	0.66
FSKF SE	0.03	0.02	0.05	0.04	0.03	0.06	0.03	0.04	0.03	0.03	0.03	0.03
NoDA MSE	0.81	0.84	0.84	0.81	0.82	1.96	0.84	1.01	0.81	0.81	0.81	0.81
NoDA SE	0.04	0.03	0.06	0.05	0.04	0.11	0.04	0.05	0.04	0.04	0.04	0.04
IndKF MSE	0.67	0.72	0.68	0.68	0.67	0.87	0.65	0.74	0.67	0.68	0.67	0.67
IndKF SE	0.03	0.02	0.05	0.04	0.03	0.06	0.03	0.04	0.03	0.03	0.03	0.03
KnRho MSE	0.79	0.85	0.81	0.80	0.79	0.99	0.74	0.87	0.79	0.79	0.79	0.79
KnRho SE	0.04	0.03	0.05	0.05	0.04	0.07	0.04	0.04	0.04	0.04	0.04	0.04
WinKF MSE	0.66	0.70	0.67	0.67	0.67	0.87	0.65	0.74	0.66	0.68	0.65	0.66
WinKF SE	0.03	0.02	0.05	0.04	0.03	0.06	0.03	0.04	0.03	0.03	0.03	0.03
New FSKF MSE	0.71	0.77	0.70	0.75	0.69	0.95	0.71	0.81	0.71	0.72	0.70	0.71
New FSKF SE	0.03	0.03	0.05	0.05	0.03	0.06	0.03	0.04	0.04	0.04	0.03	0.04
New NoDA MSE	0.81	0.85	0.85	0.81	0.82	2.01	0.84	0.99	0.81	0.81	0.81	0.81
New NoDA SE	0.04	0.03	0.06	0.05	0.04	0.11	0.04	0.06	0.04	0.04	0.04	0.04
New IndKF MSE	0.75	0.85	0.72	0.84	0.71	0.99	0.75	0.80	0.76	0.77	0.75	0.76
New IndKF SE	0.04	0.03	0.05	0.05	0.03	0.06	0.03	0.04	0.04	0.04	0.03	0.04
New KnRho MSE	0.84	0.92	0.85	0.90	0.81	1.08	0.81	0.90	0.84	0.84	0.83	0.85
New KnRho SE	0.04	0.03	0.06	0.06	0.04	0.07	0.04	0.04	0.04	0.04	0.04	0.04
New WinKF MSE	0.74	0.82	0.72	0.83	0.71	1.07	0.74	0.83	0.74	0.74	0.73	0.75
New WinKF SE	0.03	0.03	0.05	0.05	0.03	0.06	0.03	0.04	0.03	0.04	0.03	0.04
Sim Time Mean	4.41	5.06	5.75	5.62	5.20	4.96	4.81	5.14	5.48	5.11	5.04	5.13
Sim Time SE	0.58	0.67	0.84	0.82	0.69	0.66	0.64	0.68	0.80	0.70	0.67	0.68