Available online at www.sciencedirect.com
ScienceDirect
Procedia Computer Science 00 (2010) 1–9
Procedia Computer Science

International Conference on Computational Science, ICCS 2010

Data driven computing by the morphing fast Fourier transform ensemble Kalman filter in epidemic spread simulations

Jan Mandel[a,*], Jonathan D. Beezley[a], Loren Cobb[a], Ashok Krishnamurthy[a]

[a]*Department of Mathematical and Statistical Sciences, University of Colorado Denver, Denver, CO 80217-3364, USA*

Abstract

The FFT EnKF data assimilation method is proposed and applied to a stochastic cell simulation of an epidemic, based on the S-I-R spread model. The FFT EnKF combines spatial statistics and ensemble filtering methodologies into a localized and computationally inexpensive version of EnKF with a very small ensemble, and it is further combined with the morphing EnKF to assimilate changes in the position of the epidemic.

Keywords: Data assimilation, FFT, EnKF, Epidemic spread, Cell model, Covariogram
2010 MSC: 65C05, 62L12, 60G35

1. Introduction

Starting a model from initial conditions and then waiting for the result is rarely satisfactory. The model is generally incorrect, data is burdened with errors, and new data comes in that needs to be accounted for. This is a well-known problem in weather forecasting, and techniques to incorporate new data by sequential statistical estimation are known as data assimilation [1]. The ensemble Kalman filter (EnKF) [2] is a popular data assimilation method, which is easy to implement without any change in the model. The EnKF evolves an ensemble of simulations, and the model only needs to be capable of exporting its state and restarting from the state modified by the EnKF. However, the ensemble size required can be large (easily in the hundreds), the amount of computations in the EnKF can be significant, special localization techniques need to be employed to suppress spurious long-range correlations in the ensemble covariance matrix, and the EnKF does not work well for problems with sharp coherent features, such as the travelling waves found in epidemics and wildfires.

We propose a variant of EnKF based on the Fast Fourier transform (FFT), which reduces significantly the amount of computations required by the EnKF, as well as the ensemble size. The use of FFT is inspired by spatial statistic: FFT EnKF assumes that the state approximately a stationary random field, that is, the covariance between two points is mainly a function of their distance vector. Then the multiplication of the covariance matrix and a vector is a

*Corresponding author
Email address: Jan.Mandel@gmail.com (Jan Mandel)

convolution. In addition, the morphing transform [3] is used here so that changes of the state both in position and in amplitude are possible.

The FFT EnKF with morphing is illustrated here for tracking a simulated epidemic wave. The use of data assimilation techniques can increase the accuracy and reliability of epidemic tracking by using the data as soon as they are available, and some applications of data assimilation in epidemiology already exist [4, 5]. The FFT EnKF with morphing has the potential to reduce complicated simulations and accurate real-time use of data to a laptop or a smartphone in the field.

For FFT EnKF in a wildfire simulation, see [6]. The Fourier domain Kalman filter (FDKF) [7] consists of the Kalman filter used in each Fourier mode separately.

The covariance of a stationary random field can be estimated from a single realization by the covariogram [8], which can be computed efficiently by the FFT [9]. We propose to use the covariogram for an *EnKF with an ensemble of one*, which will be further developed elsewhere.

2. FFT EnKF

The EnKF approximates the probability distribution of the model state u by an ensemble of simulations $u_1, \ldots, u_N$. Each member is advanced by the simulation in time independently. When new data d arrives, it is given as data likelihood $d \sim N(Hu, R)$, where H is the *observation operator* and R is the *data error covariance matrix*. Now the *forecast ensemble* $[u_k]$ is combined with the data by the EnKF analysis [10]

$$u_k^a = u_k + C_N H^T \left(HC_N H^T + R\right)^{-1} \left(d + e_k - Hu_k^f\right), \quad k = 1, \ldots, N, \tag{1}$$

to yield the *analysis ensemble* $\left[u_k^a\right]$. Here, C_N is an approximation of the covariance C of the model state, taken to be the covariance of the ensemble, and e_k is sampled from $N(0, R)$. The analysis ensemble is then advanced by the simulations in time again. In [11], it was proved that the ensemble converges for large N to a sample from the Kalman filtering distribution when all probability distributions are Gaussian. Of course, the EnKF is used for more general cases as well.

When C_N is the ensemble covariance, the EnKF formulation (1) does not take advantage of any special structure of the model. This allows a simple and efficient implementation [12], but large ensembles, often over 100, are needed [2]. In an application, variables in the state are *random fields*, and the covariance decays with spatial distance [8]. *Tapering* is the multiplication of sample covariance term-by-term with a fixed decay function that drops off with the distance. Tapering improves the accuracy of the approximate covariance for small ensembles [13], but it makes the implementation of (1) more expensive: the sample covariance matrix can no longer be efficiently represented as the product of two much smaller dense matrices, but it needs to be manipulated as a large, albeit sparse, matrix. Random fields in geostatistics are often assumed to be stationary, that is, the covariance between two points depends on their spatial distance vector only.

The FFT EnKF discussed here uses a very small ensemble, but larger than one. We explain the FFT EnKF in the 1D case; higher-dimensional cases are exactly the same. Consider first the case when the model state consists of one block only. Denote by $u(x_i)$, $i = 1, \ldots, n$ the entry of vector u corresponding to node x_i. If the random field is stationary, the covariance matrix satisfies $C\left(x_i, x_j\right) = c\left(x_i - x_j\right)$ for some covariance function c, and multiplication by C is the convolution

$$v(x_i) = \sum_{j=1}^{n} C\left(x_i, x_j\right) u\left(x_j\right) = \sum_{j=1}^{n} u\left(x_j\right) c\left(x_i - x_j\right), \quad i = 1, \ldots, n.$$

After FFT, convolution becomes entry-by-entry multiplication of vectors, that is, multiplication by a diagonal matrix.

We assume that the random field is approximately stationary, so we neglect the off-diagonal terms of the covariance matrix in the frequency domain, which leads to the the following FFT EnKF method. First apply FFT to each member, $\widehat{u_k} = Fu_k$. Next, approximate the forecast covariance matrix in the frequency domain by the diagonal matrix with the diagonal entries given by

$$\widehat{c_i} = \frac{1}{N-1} \sum_{k=1}^{N} \left|\widehat{u_{ik}} - \overline{\widehat{u_i}}\right|^2, \quad \text{where} \quad \overline{\widehat{u_i}} = \frac{1}{N} \sum_{k=1}^{N} \widehat{u_{ik}}. \tag{2}$$

Then define approximate covariance matrix C_N by term-by-term multiplication $\cdot$ in the Fourier domain

$$u = C_N v \iff \widehat{u} = Fu, \quad \widehat{v} = \widehat{c} \bullet \widehat{u}, \quad v = F^{-1}\widehat{v}, \quad (\widehat{c} \bullet \widehat{u})_i = \widehat{c_i}\widehat{u_i}.$$

When $H = I$ and $R = rI$, the evaluation of (1) reduces to

$$\widehat{u}_k^a = \widehat{u}_k + \widehat{c} \bullet (\widehat{c} + r)^{-1} \bullet \left(\widehat{d} + \widehat{e}_k - \widehat{u}_k^f\right). \tag{3}$$

In general, the state has more than one variable, and u, C, and H have the block form

$$u = \begin{bmatrix} u^{(1)} \\ \vdots \\ u^{(n)} \end{bmatrix}, \quad C = \begin{bmatrix} C^{(11)} & \cdots & C^{(1M)} \\ \vdots & \ddots & \vdots \\ C^{(M1)} & \cdots & C^{(MM)} \end{bmatrix}, \quad H = \begin{bmatrix} H^{(1)} & \cdots & H^{(M)} \end{bmatrix}. \tag{4}$$

Here, the first variable is observed, so $H^{(1)} = I$, $H^{(2)} = 0,\ldots, H^{(M)} = 0$, and (1) becomes

$$u_k^{(j),a} = u_k^{(j)} + C_N^{(j1)} \left(C_N^{(11)} + R\right)^{-1} \left(d + e_k - u_k^{(1)}\right), \quad j = 1, \ldots, M, \tag{5}$$

and in the frequency domain

$$\widehat{u}_k^{(j),a} = \widehat{u}_k^{(j)} + \widehat{c}^{(j1)} \bullet \left(\widehat{c}^{(11)} + r\right)^{-1} \bullet \left(\widehat{d} + \widehat{e}_k - \widehat{u}_k\right). \tag{6}$$

The cross-covariance between field j and field 1 is approximated by neglecting the off-diagonal terms of the sample covariance in the frequency domain as well,

$$\widehat{c}_i^{(j1)} = \frac{1}{N-1} \sum_{k=1}^{N} \left(\widehat{u}_{ik}^{(j)} - \overline{\widehat{u}}_i^{(j)}\right)\left(\widehat{u}_{ik}^{(1)} - \overline{\widehat{u}}_i^{(1)}\right), \quad \text{where} \quad \overline{\widehat{u}}_i^{(\ell)} = \frac{1}{N} \sum_{k=1}^{N} \widehat{u}_{ik}^{(\ell)}, \ \ell = 1, j. \tag{7}$$

In the computations reported here, we have used the real sine transform, so all numbers in (7) are real. Also, the use of the sine transform naturally imposes no change of the state on the boundary.

3. Morphing EnKF

Given an initial state u, the initial ensemble in the morphing EnKF [3, 12] is given by

$$u_k^{(i)} = \left(u_{N+1}^{(i)} + r_k^{(i)}\right) \circ (I + T_k), \quad k = 1, \ldots, N, \tag{8}$$

with an additional member $u_{N+1} = u$, called the *reference member*. In (8), $r_k^{(i)}$ are random smooth functions on Ω, T_k are random smooth mappings $T_k : \Omega \to \Omega$, and $\circ$ denotes composition. Thus, the initial ensemble varies both in amplitude and in position, and the change position is the same in all blocks. The random smooth functions and mapping are generated by FFT as Fourier series with random coefficients with zero mean and variance that decays quickly with frequency.

The data d is an observation of $u^{(1)}$. The first blocks of all members $u_1, \ldots, u_N$ and d are then registered against the first block of u_{N+1} as

$$u_k^{(1)} \approx u_{N+1}^{(1)} \circ (I + T_k), \quad T_k \approx 0, \quad \nabla T_k \approx 0, \quad k = 0, \ldots, N,$$

$u_0^{(1)} = d$ and $T_k : \Omega \to \Omega$, $k = 0, \ldots, N$ are called *registration mappings*. The registration mapping is found by multilevel optimization. The *morphing transform* maps each ensemble member u_k into the extended state vector, the *morphing representation,*

$$u_k \mapsto \widetilde{u}_k = M_{u_{N+1}}(u_k) = \left(T_k, r_k^{(1)}, \ldots, r_k^{(M)}\right), \tag{9}$$

where $r_k^{(j)} = u_k^{(j)} \circ (I + T_k)^{-1} - u_{N+1}^{(j)}$, $k = 0, \ldots, N$, are *registration residuals*. Likewise, the extended data vector is defined by $d \mapsto \widetilde{d} = \left(T_0, r_0^{(1)}\right)$ and the the observation operator is $\left(T, r^{(1)}, \ldots, r^{(M)}\right) \mapsto \left(T, r^{(1)}\right)$. We then apply the

FFT EnKF method (6) is applied to the transformed ensemble $\widetilde{u}_1, \ldots, \widetilde{u}_N$. The covariance $C^{(11)}$ in (5) consists of three diagonal matrices and we neglect the off-diagonal blocks, so the fast formula (6) can be used. The analysis ensemble $u_1, \ldots, u_{N+1}$ is obtained by the *inverse morphing transform*

$$u_k^{a,(i)} = M_{u_{N+1}}^{-1}\left(\widetilde{u}_k^a\right) = \left(u_{N+1}^{(i)} + r_k^{a,(i)}\right) \circ \left(I + T_k^a\right), \quad k = 1, \ldots, N+1, \tag{10}$$

where the new transformed reference member is given by

$$\widetilde{u}_{N+1}^a = \frac{1}{N}\sum_{k=1}^{N} \widetilde{u}_k^a. \tag{11}$$

4. Epidemic model

The epidemic model that we used for this study is a spatial version of the common S-I-R dynamic epidemic model. A person is *susceptible* or *infectious* in this context if he or she can contract or transmit the disease, respectively. The *removed* state includes those who have either died, have been quarantined, or have recovered from the disease and become immune. The state variables are the susceptible (S), the infectious (I), and the removed (R) population densities.The core ideas for this model date back to the 1957 spatial formulation by Bailey [14], but the specific version that we have employed here is due to Hoppenstaedt [15, p. 64].

The population is considered to be dispersed over a planar domain $\Omega \subset \mathbb{R}^2$, and it is labelled according to its position with respect to the spatial coordinates x and y. The (deterministic) evolution of the state $(S(t), I(t), R(t))$ is given by

$$\left.\begin{array}{l} \frac{\partial S(x,y,t)}{\partial t} = -S(x,y,t) \iint_{\Omega} w(x,y,u,v)\, I(u,v,t)\, dudv, \\ \frac{\partial I(x,y,t)}{\partial t} = S(x,y,t) \iint_{\Omega} w(x,y,u,v)\, I(u,v,t)\, dudv - q(x,y,t)\, I(x,y,t), \\ \frac{\partial R(x,y,t)}{\partial t} = q_i(x,y,t)\, I(x,y,t). \end{array}\right\} \tag{12}$$

The function $q(x,y,t)$ gives the rate of removal of infectives due to death, quarantine, or recovery. The weight function $w(x,y,u,v)$ measures the influence of infectives at spatial position (u,v) on the exposure of susceptibles at position (x,y); in this simulation we used the function $w(x,y,u,v) = \alpha \exp[-((x-u)^2 + (y-v)^2)^{1/2}/\lambda]$, which expresses the idea that the influence of nearby infectives decays as an exponential function of Euclidean distance, with constant λ, characteristic of the distance at which the disease spreads. More mobile societies will have larger values of λ. The parameter α measures the infectiousness of the disease.

A stochastic cell model is created by treating the quantities on the right-hand-side of (12) as the intensities of a Poisson process and by piecewise constant integration over the cells. The domain Ω is decomposed into nonoverlapping cells Ω_i with centers (x_i, y_i) and areas $A(\Omega_i)$, $i = 1, \ldots, K$. The state in the cell Ω_i is the random element (S_i, I_i, R_i), advanced in time over the interval $[t, t+\Delta t]$ by

$$S_i(t+\Delta t) = S_i(t) - \Delta S_i, \quad I_i(t+\Delta t) = I_i(t) + \Delta S_i - \Delta R_i, \quad R_i(t+\Delta t) = R_i(t) + \Delta R_i,$$

where the random increments ΔS_i and ΔR_i are sampled from

$$\begin{array}{l} \Delta S_i \sim \mathrm{Pois}\left(S_i(t) \sum_{j=1}^{K} w\left(x_i, y_i, x_j, y_j\right) I_j(t)\, A(\Omega_i)\, \Delta t\right), \\ \Delta R_i \sim \mathrm{Pois}\left(q_i(t)\, I_i(t)\, A(\Omega_i)\, \Delta t\right), \end{array} \tag{13}$$

and $q_i(t)$ is the given removal rate in the cell Ω_i. The summation in (13) is done only over the cells Ω_j near Ω_i; for far away cells, the weights $w\left(x_i, y_i, x_j, y_j\right)$ are negligible. It is not necessary to compute a Poisson-distributed transmission rate from each source cell to a given target cell, because a finite sum of independent Poisson-distributed random variables, each with its own intensity parameter, is itself Poisson-distributed with an intensity parameter equal to the sum of the individual intensities.

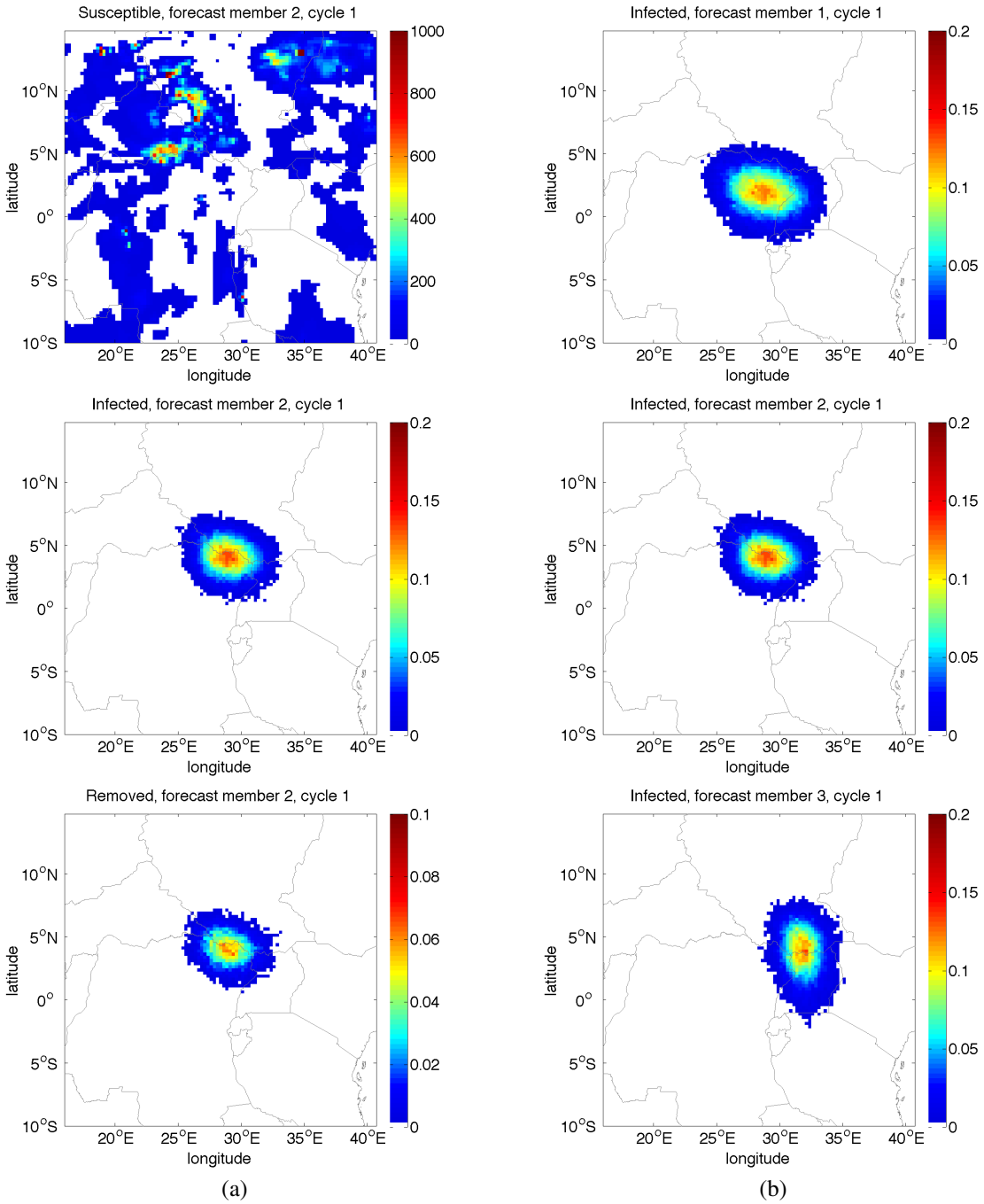

Figure 1: (a) The number of people per kilometer squared infected, susceptible, and removed after 120 time steps in a simulation of an epidemic disease spreading through central Africa. These images correspond to variables I, S, and R in Equation (12). (b) Number of people infected per kilometer squared in three forecast ensemble members.

5. Computational results

We have chosen to model an epidemic disease that first emerges in Congo. The computational domain is a square portion of central Africa. In Figure 1 (a), we see the epidemic wave 120 model time steps after the emergence of the disease. The behavior of the model is such that any spurious infection will tend to grow into a secondary infection wave. This is problematic for data assimilation because the occurrence of spurious features is virtually guaranteed. We attempt to reduce the occurrence and magnitude of these features using the morphing transformation and FFT EnKF; however, some amount of residual artifacts will remain. We have found that by processing the model state in the following manner, we can further reduce these artifacts. We begin by scaling the absolute quantities contained in the model variables to a percentage of the local population before performing the data assimilation. After data assimilation, we truncate the variables to the range $[0, 1]$, and we apply a threshold so that any infection rate below 1% is set to 0. Finally, we rescale the output in absolute units ensuring that the number of people at each grid cell is preserved. We have applied the FFT EnKF to the epidemic model described in Section 4 with an ensemble of size 5. Each ensemble simulation was started with the same initial conditions, but with different random seeds, and advanced in time by 100 model time units, then perturbed randomly to obtain the initial ensemble. The analysis ensemble and data were advanced in time an additional 20 model time steps for further assimilation cycles. In total, 3 assimilation cycles were performed in this manner.

We have perturbed each member of the initial ensemble randomly in space by applying (10) to the each variable of the morphing representation of the model. The mappings T_k for this perturbation were generated from a space of smooth functions that are zero at the boundary. While the residuals r_k are customarily initialized to smooth random fields as well, we have chosen to set $r_k = 0$ to avoid spurious infections. We instead multiply each field after the inverse morphing transform by $1 + s_k$, where s_k is another smooth random field. This ensures that an initial infection rate of 0 is unchanged by the perturbation. A part of a typical ensemble with spatial as well as amplitude variability is shown Figure 1 (b).

The output of the observation function used in this example consists of the *Infected* field of the model. In this case, the data is a spatial "image" of the number of infected persons in each grid cell. The data were generated synthetically from a model simulation, which was initialized in the same manner as the ensemble.

Four variants of the EnKF were then applied: the standard EnKF and FFT EnKF and morphing EnKF and morphing FFT EnKF. The same initial ensemble and the same data were used for each method. The deviation of the initial ensemble and the model error were chosen so that the analysis should be about half way between the forecast and the data. In the morphing variants, the data deviation in the amplitude was taken very large, so that the filter updates essentially only the position. Ensemble of size 5 was used. The result in the first assimilation cycle for each method is shown in Figures 2 and 3. The first image in each column is the forecast mean. In the morphing variants, the mean is taken over all ensemble members in all fields of the morphing representation (9) and it plays the role of the comparison state for registration. Thus, in the morphing variants, both the amplitude and the position of the infection wave in the ensemble members are averaged. The second image in each column is the data, which is a model trajectory started from the same initial state for each method. Because the model is itself stochastic, the data images are slightly different. The third image in each column is the analysis mean, which is taken in the morphing representation (11) for two morphing filters, so that both the amplitude and the location are averaged.

We see that both standard EnKF and FFT EnKF filters cannot move the state towards the data; a much larger ensemble would be needed. The morphing EnKF does move the state towards the data, but there are strong artifacts due to the poor approximation of the covariance by the covariance of the small ensemble. Finally, the morphing FFT-EnKF is capable of moving the state towards the data better.

6. Conclusion

We have introduced morphing FFT EnKF and presented preliminary results on data assimilation for an epidemic simulation. Morphing was essential to move the state towards the data, but it resulted in artifacts for the small ensemble size used, yet small ensemble size is important to perform simulations with data assimilation on general computing devices instead of supercomputers. We have observed that the estimation of the covariance matrix in the

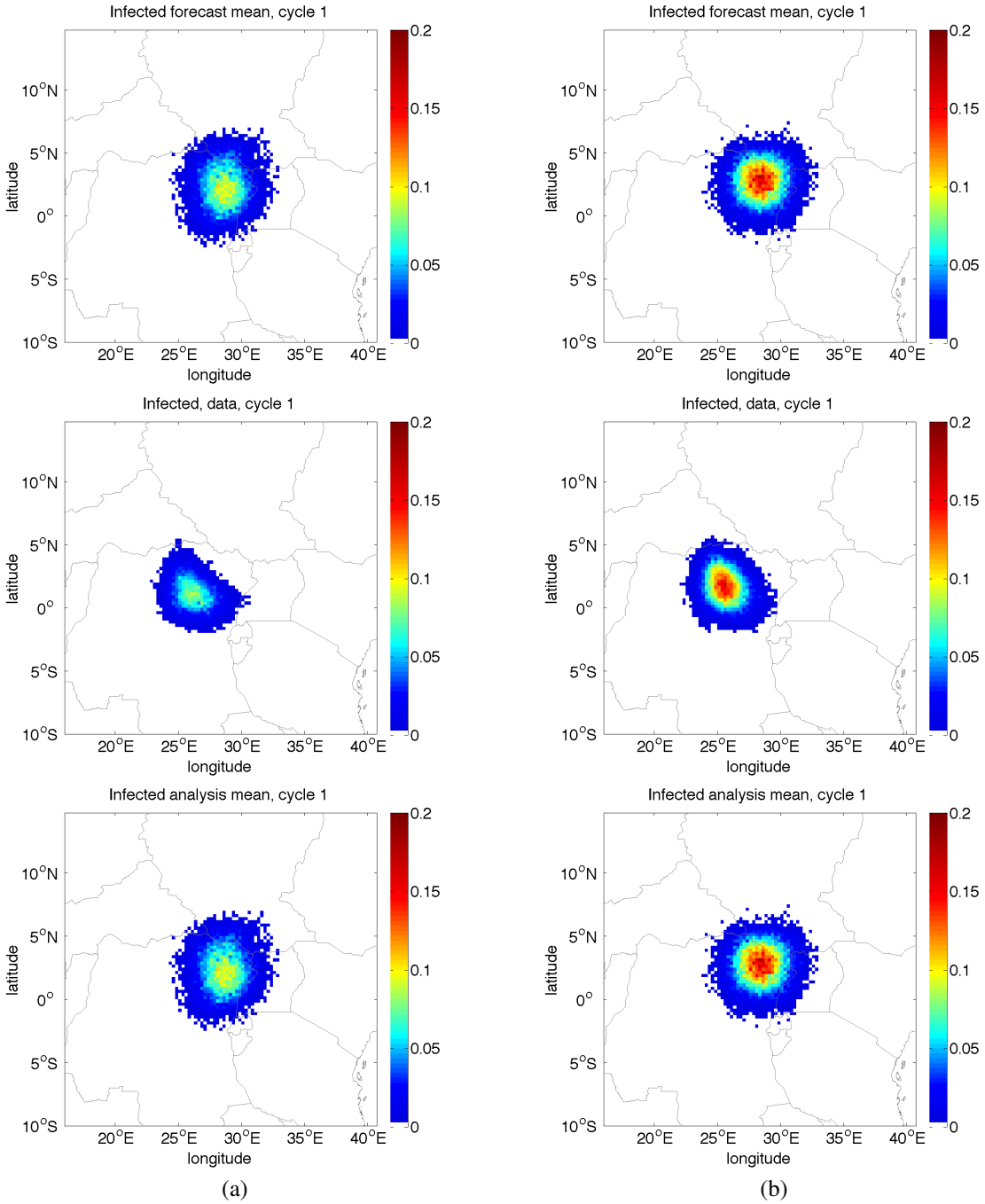

Figure 2: The number of people infected per kilometer squared in analysis cycle 1 using the standard EnKF and FFT EnKF, each with ensemble size of 5. Both approaches are unable to move the location of the infection in the simulation state.

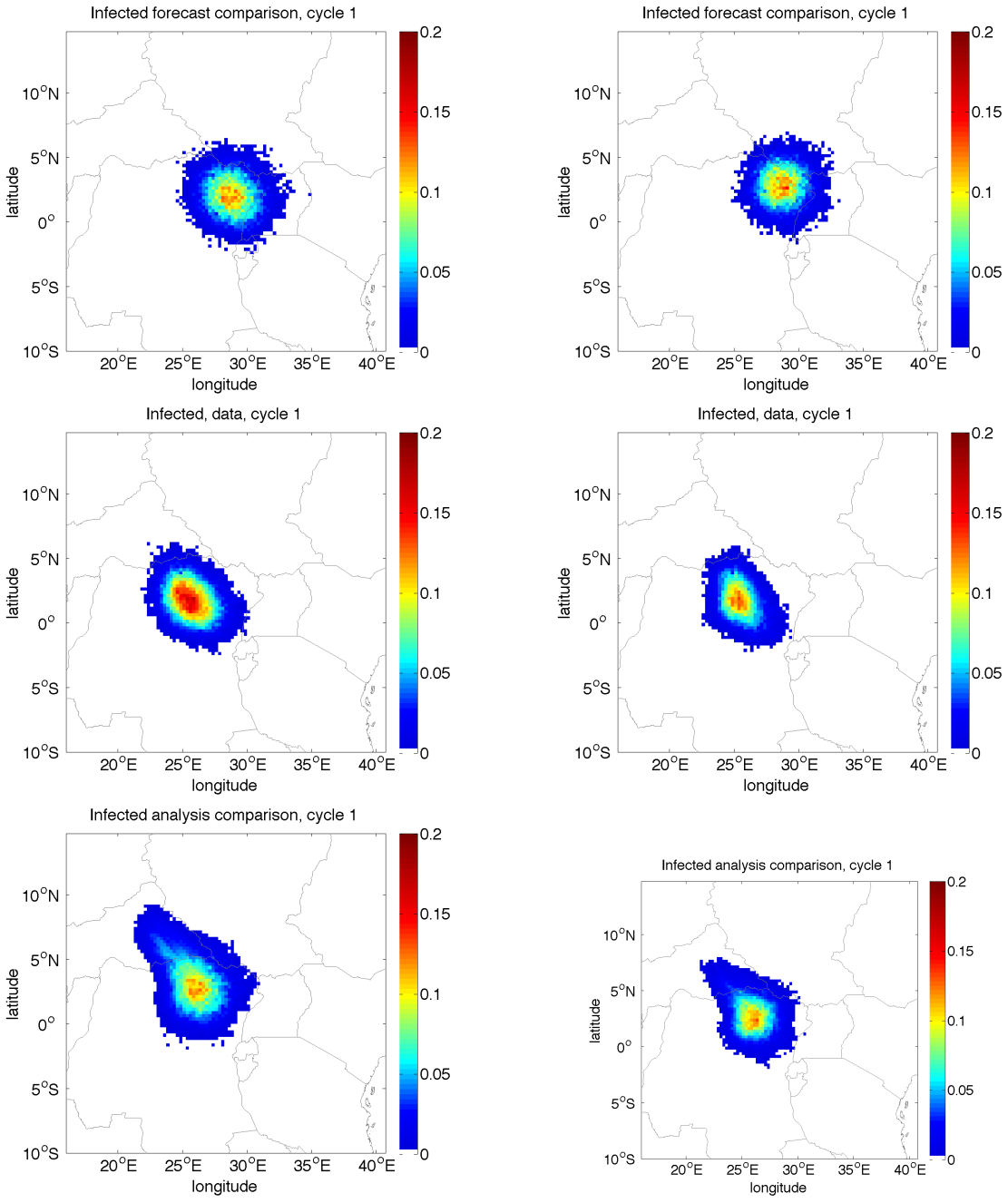

Figure 3: The number of people infected per kilometer squared in analysis cycle 1 using the morphing EnKF and morphing FFT EnKF, each with ensemble size of 5. Both approaches are able to move the state spatially and perform similarly. However, EnKF suffers from stronger artifacts due to low accuracy and low rank of the ensemble covariance than the morphing FFT EnKF.

frequency domain results in better forecast covariance in the algorithm, which has the potential to reduce the artifacts due to small ensemble size.

7. Acknowledgements

This work was partially supported by NIH grant 1 RC1 LM01641-01 and NSF grants CNS-0719641 and ATM-0835579.

References

[1] E. Kalnay, Atmospheric Modeling, Data Assimilation and Predictability, Cambridge University Press, 2003.
[2] G. Evensen, Data Assimilation: The Ensemble Kalman Filter, 2nd Edition, Springer Verlag, 2009.
[3] J. D. Beezley, J. Mandel, Morphing ensemble Kalman filters, Tellus 60A (2008) 131–140.
[4] L. Bettencourt, R. Ribeiro, G. Chowell, T. Lant, C. Castillo-Chavez, Towards real time epidemiology: data assimilation, modeling and anomaly detection of health surveillance data streams, in: Intelligence and Security Informatics: Biosurveillance, Vol. 4506 of Lecture Notes in Computer Science, Springer, 2007, pp. 79–90.
[5] C. Rhodes, T. Hollingsworth, Variational data assimilation with epidemic models, Journal of Theoretical Biology 258 (4) (2009) 591–602.
[6] J. Mandel, J. D. Beezley, V. Y. Kondratenko, Fast Fourier transform ensemble Kalman filter with application to a coupled atmosphere-wildland fire model, arXiv:1001.1588, International Conference on Modeling and Simulation (MS'2010), accepted (2010).
[7] E. Castronovo, J. Harlim, A. J. Majda, Mathematical test criteria for filtering complex systems: plentiful observations, J. Comput. Phys. 227 (7) (2008) 3678–3714.
[8] N. A. C. Cressie, Statistics for Spatial Data, John Wiley & Sons Inc., New York, 1993.
[9] D. Marcotte, Fast variogram computation with FFT, Computers & Geosciences 22 (10) (1996) 1175–1186.
[10] G. Burgers, P. J. van Leeuwen, G. Evensen, Analysis scheme in the ensemble Kalman filter, Monthly Weather Review 126 (1998) 1719–1724.
[11] J. Mandel, L. Cobb, J. D. Beezley, On the convergence of the ensemble Kalman filter, arXiv:0901.2951, Applications of Mathematics, to appear (January 2009).
[12] J. Mandel, J. D. Beezley, J. L. Coen, M. Kim, Data assimilation for wildland fires: Ensemble Kalman filters in coupled atmosphere-surface models, IEEE Control Systems Magazine 29 (2009) 47–65.
[13] R. Furrer, T. Bengtsson, Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants, J. Multivariate Anal. 98 (2) (2007) 227–255.
[14] N. Bailey, Mathematical Theory of Epidemics, Griffin, 1957.
[15] F. Hoppenstaedt, Mathematical Theories of Populations, Demographics, and Epidemics, CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM, 1975.