Ensemble-based data assimilation and the localisation problem

Ruth E. Petrie and Sarah L. Dance
University of Reading

The *butterfly effect* is a popularly known paradigm; commonly it is said that when a butterfly flaps its wings in Brazil, it may cause a tornado in Texas. This essentially describes how weather forecasts can be extremely sensitive to small changes in the given atmospheric data, or initial conditions, used in computer model simulations. In 1961, Edward Lorenz found that small changes in initial conditions given to a weather forecast model can, in time, lead to entirely different forecasts (Lorenz, 1963). This discovery highlights one of the major challenges in modern weather forecasting: to provide the computer model with the most accurately specified initial conditions possible. A process known as *data assimilation* seeks to minimize the errors in the given initial conditions and in 1911 was described by Bjerknes as *the ultimate problem in meteorology* (Bjerknes, 1911).

Weather forecast models

Weather forecasts produced by institutions such as the UK's Met Office are generated by sophisticated Numerical Weather Prediction (NWP) computer models. These models take given initial conditions of the atmospheric state and evolve them in time according to the relevant governing dynamical laws to produce a weather forecast. Initial conditions of atmospheric variables such as temperature, wind speed and direction, pressure, and humidity need to be known to describe the current atmospheric state. NWP models require information on each atmospheric variable for each grid box in the model, i.e. currently a total of around 10^7 variables, to fully describe the initial conditions.

Traditionally, measurements are taken by instrumentation on radiosondes, ships, aeroplanes and at surface stations. The data collected by these instruments are distributed sparsely over the globe. The Southern Hemisphere and Polar regions are deficient in data generated by these traditional methods due to their geographical nature. Observational data from satellites has been available since the 1960s. Satellites take measurements of radiance from which atmospheric data, such as temperature, are inferred; these are known as *indirect observations*. Satellites provide almost global data coverage and this has made it possible to gain information in remote locations where data had been previously unavailable. Data from satellites now constitute the majority of observational data used in modern weather forecasting (Simmons, 2003). Despite all the observational data available, currently around 10^6 variables, they are still insufficient to fully specify the initial conditions. It is therefore impossible to rely only on these observational data to supply the NWP model with an initial state of the atmosphere. It is possible to use data from the NWP model itself, following a spin-up period, to provide a complete set of initial conditions; however, these are unlikely to be accurate enough on their own to provide meaningful forecasts.

Data assimilation

In order to provide the most accurate possible initial conditions, observational data are combined with a forecast estimate of the atmospheric state using sophisticated mathematical techniques; this process is known as *data assimilation*. The forecast estimate is generally a short-term forecast from a previous model run and it is referred to as the *background state* of the atmosphere. The background state is an approximate representation of the atmospheric state; therefore there exists a degree of uncertainty in this estimate. Similarly the observations are not perfect and have an associated uncertainty. Mathematical formulae are used to combine the observational data with the background state, according to a weighting of the uncertainty in the background state and the observational data. This combined state is known as the *analysis state*. This should be a more accurate representation of the *true* state of the atmosphere than the background state and will provide complete initial conditions for the NWP model. The aim of data assimilation is to produce the most accurate analysis estimate possible of all the atmospheric variables at every grid box required by the NWP model given the background state, observations and their associated errors.

The specification of the uncertainty or error statistics in both the observations and background state is of crucial importance in data assimilation. These errors are represented in *covariance matrices*. A covariance matrix represents correlations between state elements. If two atmospheric state variables are closely related (i.e. they vary together, such as wind and pressure due to geostrophic balance) then they will have a high correlation; this is represented by an element in the matrix. The background error covariance matrix, denoted **P,** also describes the degree of confidence in the background state through the diagonal elements of **P**, which are known as *variances*. If the diagonal elements of **P** are small, the associated error is small and there is high confidence that the background state estimate is a good one. Conversely, if the diagonal elements of **P** are large then there is little confidence in the quality of the background state estimate. Similarly there exists an observational error covariance matrix, **R**, which describes the confidence in the accuracy of the observations. During the data assimilation process, a weighted average of the background covariance, **P**, and the observational covariance, **R**, is calculated. This weighting determines to what degree the previous forecast state can be adjusted by the assimilation of the observations. If either one of the covariance matrices is incorrectly specified, this weighting will also be incorrectly specified, the analysis state may be less accurate and there is potential for the forecast to be degraded. It is currently a major challenge in meteorology to specify these matrices to describe the true uncertainties and correlations as accurately as possible.

The Kalman filter

The Kalman filter (Kalman and Bucy, 1961) is a well established form of sequential data assimilation. Filter is a technical term for a data assimilation scheme that uses only observational data valid up to and including the analysis time. The Kalman filter operates by updating the forecast trajectory, at each observation time, by explicitly solving

RMetS Royal Meteorological Society

a series of equations (a schematic is shown in Figure 1). There are two stages involved in using the Kalman filter equations to assimilate observational data to reach the analysis state of the system: the forecast stage and the analysis stage.

In the forecast stage, the background state, •, is forecast by the dynamical system equations to the time of the first observation. The error statistics in the background error covariance matrix **P**, illustrated by the grey shading in Figure 1, are also evolved in time by the model dynamics. This allows for an accurate representation of the error statistics throughout the assimilation.

In the analysis stage, a weighting between the observational and background errors is calculated as part of the Kalman filter equations (Kalman and Bucy, 1961). This error weighting determines the degree to which the forecast trajectory can be adjusted towards the observational data. If the error statistics of the background state estimate are low, then there is high confidence that the estimate is a good one and the assimilation of the observational data will have little impact. Conversely, if the background error statistics are large then there is little confidence in the quality of the forecast estimate and the assimilation of the observational data will have a much greater impact on the forecast trajectory. The analysis state, •, is used to make a forecast of this estimate to the time of the next observation which then becomes the background state for the next assimilation.

This system allows an atmospheric state and its error covariance to evolve in accordance with the dynamical equations when there are no observations but allows the forecast to be 'corrected' when observations become available.

The standard Kalman filter is impractical for implementation in operational NWP partly due to the computational expense of calculating and evolving the background error covariance matrix. Variations have been developed from the basis of the standard Kalman filter equations. One such variation is the ensemble Kalman filter (EnKF) (Evensen, 2003). This replaces the single forecast trajectory of the standard Kalman filter with an ensemble.

The initial ensemble is a collection of perturbed state estimates of the background state. The EnKF produces an ensemble of forecasts from which an approximation to the evolved background error covariance can be calculated. This approximation provides the EnKF with significant computational savings compared to the standard Kalman filter. The spread of the ensemble allows the forecast to be qualified by a degree of uncertainty. This is illustrated in Figure 2, where the degree of uncertainty is supposed to mimic the grey shading of Figure 1. If all the members of the ensemble predict a similar state, i.e. are tightly spaced, then there is high confidence in the analysis; conversely if the ensemble members are spread widely, then there is low confidence in the analysis state. After the assimilation of observational data, the spread of the ensemble members is reduced. Another key benefit of the EnKF over the standard Kalman filter is that the ensemble of analysis states could be used to provide good initial perturbations for an ensemble-based NWP scheme.

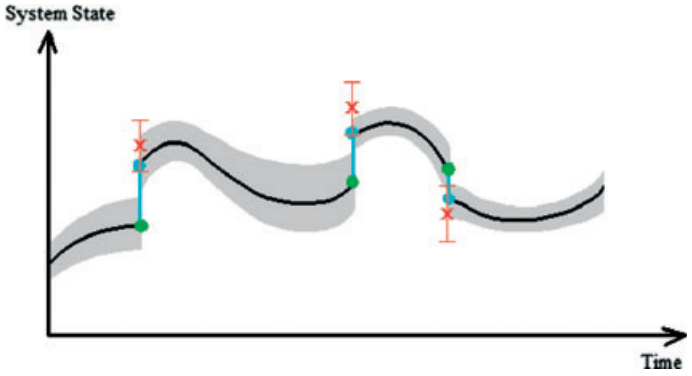

Figure 1. Schematic of a Kalman Filter. The green dots indicate the background states at various times, the blue dots indicate the analysis states and the blue lines are the analysis increments. The black line is the forecast trajectory. The grey shading indicates the uncertainty in the background forecast and the red crosses indicate observations with error bars in red.

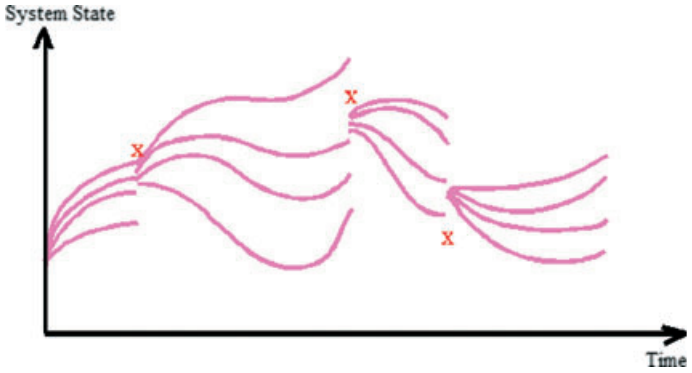

Figure 2. Schematic of an ensemble Kalman filter. The lines represent individual ensemble members, the red crosses indicate observations.

Problems associated with the ensemble Kalman filter

Maintaining ensembles of state estimates is computationally expensive. Although increases in computing power have enabled such statistical approaches to become more feasible, they are not widely implemented as the cost is often still prohibitively large. The cost is dependent on the size of the ensemble. Therefore reducing the number of ensemble members can reduce the computational cost. Care must be taken, however, to ensure that the ensemble size is not too small so that it remains statistically representative of the system (Kalnay, 2003). The number of ensemble members required to represent the system is related to the size of the state space. Current NWP models have state spaces of the order of 10^7 elements and can thus require a large ensemble to adequately represent the statistics.

Undersampling

In situations where the ensemble size is too small to be statistically representative of the system it is said to be *undersampled*. Undersampling is a fundamental problem in ensemble Kalman filtering. The success of the EnKF is highly dependent on the size of the ensemble being adequate (Houtekamer and Mitchell, 1998). Undersampling introduces three major problems in ensemble filtering: underestimation of covariance, filter divergence and the development of long-range spurious correlations, which are now discussed.

Underestimation of covariance

The evolved background error covariances are also known as the *forecast error covariances* and are systematically underestimated after each assimilation cycle (Furrer and Bengtsson, 2007). If this covariance is underestimated, false confidence is placed in the background state. The systematic underestimation is not an indication that the analysis state is really more accurately representing the *true* system state. The smaller

the ensemble is, the greater the degree of undersampling that is present and the greater the chance of underestimated forecast error covariances (Ehrendorfer, 2007). The underestimation of the forecast error covariances can lead to subsequent problems such as filter divergence (Hamill *et al.,* 2001), and the development of spurious long-range correlations (Ehrendorfer, 2007).

Filter divergence

As a filter progressively underestimates the forecast error covariances, it erroneously becomes more confident in the accuracy of the forecast state estimate, giving less weighting to the observations in the assimilation. Subsequently, the observational data are progressively ignored. This is known as *filter divergence*. It can be thought of as all of the ensemble members converging on an incorrectly specified analysis state. This can be seen in Figure 3. At the start of the assimilation, time = 0s, the ensemble spread is large but by the end of the assimilation window, 100s, the size of the ensemble spread has decreased significantly.

Spurious correlations

Observations made at one location can have an impact on state variables that are physically remote from the observation location, through the filter equations (Anderson, 2001). In the physical world it is expected that correlations with any given observation point will decrease with distance. Therefore, correlations between state components that are not physically related and that are spatially remote may be regarded as spurious and will degrade the quality of the analysis estimate. These correlations can be caused by undersampling; as the degree of undersampling decreases so the problem of spurious correlations also decreases (Lorenc, 2003).

Methods of mitigation

It is not currently known how to eliminate these problems but various methods have been developed to negate their impact. Covariance inflation and covariance localisation (Hamill *et al.,* 2001) are two of these methods and are now reviewed.

Covariance inflation

Covariance inflation, introduced by Anderson and Anderson (1999), is a method of correcting the underestimation in the forecast error covariance matrix. The principle is to simply increase the forecast error covariances by an inflationary factor, r, in each assimilation cycle to negate the systematic underestimation. The inflation factor is normally chosen, based on experience, to be slightly greater than 1.0, although it can be much greater than 1.0. This technique of covariance inflation is commonly used in ensemble filtering. Although this is a useful technique, it can lead to physical balances in the system dynamics being disrupted by the inflations (Anderson, 2001). Inflation factors do not help to correct the problem of long-range spurious correlations; for this a more sophisticated approach is required.

Covariance localisation

Covariance localisation (Houtekamer and Mitchell, 1998; Hamill *et al.,* 2001) is a process of 'cutting off' long-range spurious correlations in the error covariance matrices, thus helping to improve the estimate of the forecast error covariance. It is ordinarily achieved by first defining a correlation matrix, C, as shown in Figure 4(a), and then taking a *Schur product* (Schur, 1911) of this correlation matrix and the forecast error covariance. A Schur product involves an elementwise product of matrices and is written as C ∘ **P**. An example of a Schur product for a 2 × 2 matrix is as follows:

$$\begin{pmatrix}1 & 2\\ 3 & 4\end{pmatrix}\circ\begin{pmatrix}8 & 7\\ 6 & 5\end{pmatrix}=\begin{pmatrix}1\times 8 & 2\times 7\\ 3\times 6 & 4\times 5\end{pmatrix}=\begin{pmatrix}8 & 14\\ 18 & 20\end{pmatrix}$$

Representations of covariance matrices are shown in Figure 4. Each pixel in the grid represents a covariance such that, for example, pixel (10, 40) is a measure of covariance between the 10th and 40th state variables. The pixels are coloured according to the size of the covariance. The correlation matrix, C, is chosen such that the structure is a band of non-zero elements around the leading diagonal, with ones on the diagonal, falling to zero at a specified distance from the diagonal. Figure 4(b) shows a sample forecast error covariance with undesirable spurious correlations; these can be seen in those elements that are far from the leading diagonal and have large correlations. Figure 4(c) shows the forecast error covariance after it has been localized; correlations that existed at a large distance from the leading diagonal in Figure 4(b) have been 'cut-off' or removed while the correlations which are physically local, close to the diagonal, have been maintained. This localised matrix is more suitable for use in the assimilation process to describe the forecast uncertainty as the spurious correlations have been removed.

The distance at which correlations in the error covariance matrices are cut-off (reduced to zero) is known as the *filtering length scale*. It is essential that while unphysical, remote, spurious correlations are removed by the correlation function, correctly specified physical correlations are not excessively damped but maintained. If the filter length scale is too long, so as to allow all the dynamical correlations, then many of the spurious correlations may not be removed. If the filter length scale is too short, then important physical dynamical correlations may be lost as well as the spurious ones. It is important that this length scale is correctly chosen for a given system, though at present defining the length scale is a heuristic (experience-based) process.

One additional benefit of applying Schur product localisation is that the *effective* size of the ensemble is increased (Oke *et al.,* 2007). This has a feedback effect negating some of the problems associated with undersampling. Another benefit of this localisation is that after applying the Schur product, the covariance matrix becomes sparse, i.e. it has many zeros. This can lead to important computational savings (Lorenc, 2003).

There are drawbacks to this technique. Important information on physical dynamical relationships is held within the error covariance matrices; the modification of these matrices by the Schur product is

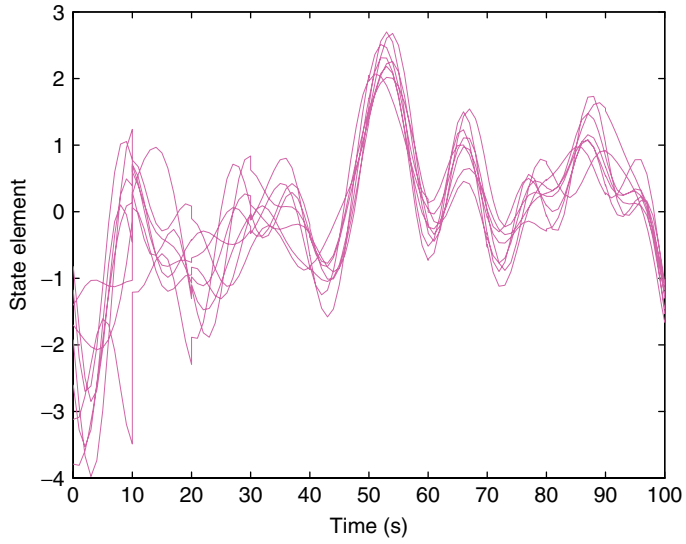

Figure 3. Each line is the forecast estimate of one ensemble member. The convergence of the ensemble, or filter divergence, due to underestimation of covariance is shown.

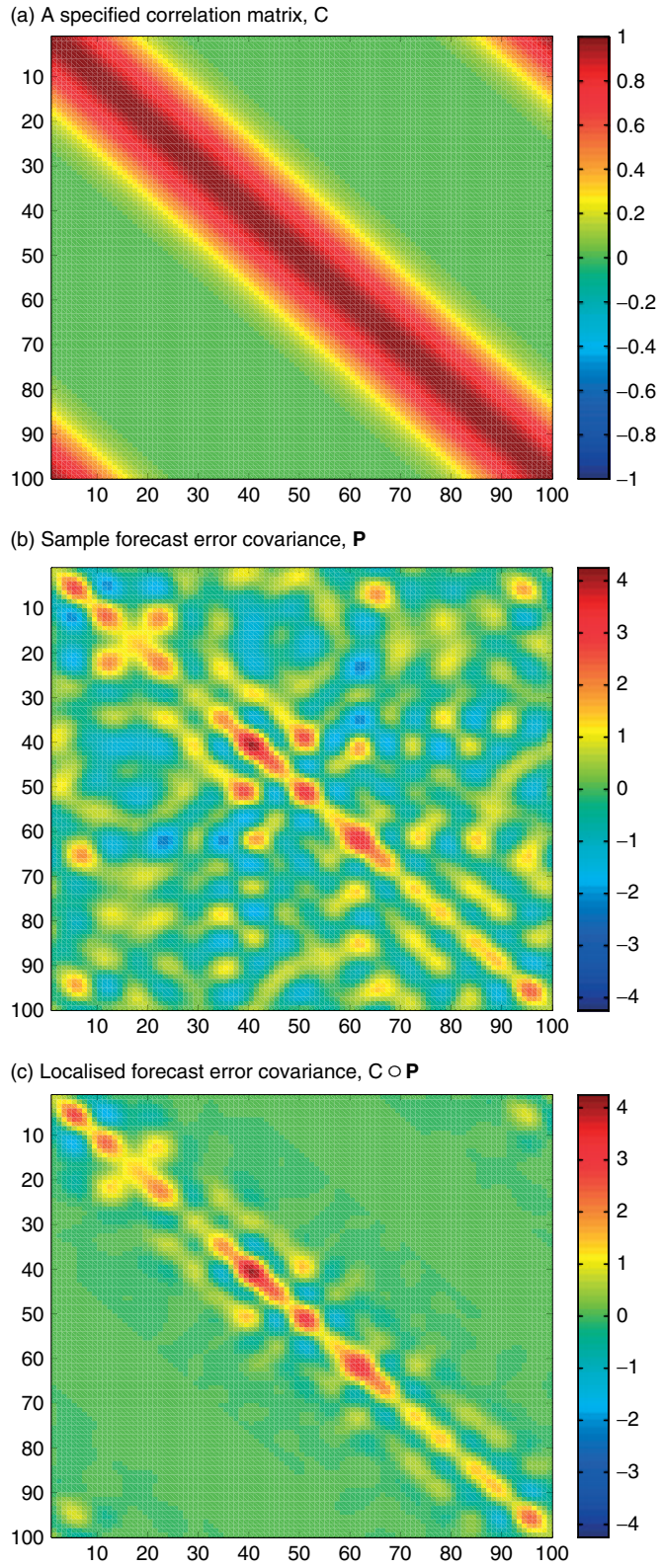

*Figure 4. Schematic representation of covariance localization with 100 state elements. Figure 4a shows an example of a correlation matrix, C, which is a commonly used mathematical function. The forecast error covariance, **P**, shown in Figure 4b, was produced using a simple advection model (Petrie, 2008), with periodic boundary conditions which can be seen in the top right and bottom left corners. Figure 4c is the Schur product composition, C○ **P** , of the correlation covariance matrix and the forecast error covariance matrix.*

known to disrupt these balances (Oke *et al.,* 2007). This is highly undesirable as it would mean that related variables would not be constrained to each other during the forecast, leading to an invalid forecast.

Square root filters

A *square root filter* is one that does not use the background error covariance matrix explicitly, rather it uses a different matrix, **X′**. This is the background error perturbation matrix and it holds only information on each ensemble member's deviation from the ensemble mean as the columns of **X′**. The perturbation matrix, **X′**, is a form of square root of the full **P** matrix. Many formulations of ensemble filters use this matrix instead of the full covariance matrix as it is computationally efficient (Bishop *et al.,* 2001). The Ensemble Transform Kalman Filter (ETKF) is an example of one type of square root filter introduced by Bishop *et al.* (2001). This is a beneficial algorithm for operational implementation as it is able to rapidly calculate the forecast error covariance and is computationally very efficient. Another benefit of the ETKF is its ability to identify an optimal site for an observation to improve a given forecast region; this is covered in detail in Bishop *et al.* (2001). The ETKF is a popular choice in ensemble forecasting and has become operational in some centres.

The use of the perturbation matrix instead of the full covariance matrix makes covariance localisation difficult. Recall that the Schur product localisation is achieved by taking a Schur product of the correlation matrix with the full covariance matrix, C ○ **P**, and not the perturbation matrix. This can be written in terms of square root matrices as:

$(\rho\,\rho^{T}) \circ (\mathbf{X'}\,\mathbf{X'}^{T})$; note ρ is the square root of the correlation matrix.

To achieve covariance localisation in a square root filter, an approximation to the Schur product localisation can be written mathematically as:

$$(\rho\,\rho^{T}) \circ (\mathbf{X'}\,\mathbf{X'}^{T}) \sim (\rho \circ \mathbf{X'})(\rho \circ \mathbf{X'})^{T}.$$

It is known that this is merely an approximation and not equality; however, it was implemented and tested within an ETKF algorithm to ascertain if this was a reasonable approximation.

To test this approximation, the covariance matrices representing the left- and right-hand sides were plotted and compared. Figure 5(a) is identical to Figure 4(c) as $(\rho\rho^{T}) \circ (\mathbf{X'}\,\mathbf{X'}^{T}) = C \circ \mathbf{P}$. Figure 5(b) shows the test of the approximation. If the approximation is a reasonable one then the Figures 5(a) and 5(b) should be similar; however, they are clearly very different. The approximation to localisation in Figure 5(b) has not achieved localisation as many elements far from the leading diagonal have not been removed. In addition, the magnitudes of the covariances have been reduced. Therefore it would be inappropriate to use this approximation to achieve covariance localisation in a square root filter. A more sophisticated approach, such as that in Bishop and Hodyss (2009), is required.

Summary

NWP models cannot rely solely on either observational data or model forecasts to fully and accurately describe the atmospheric system.

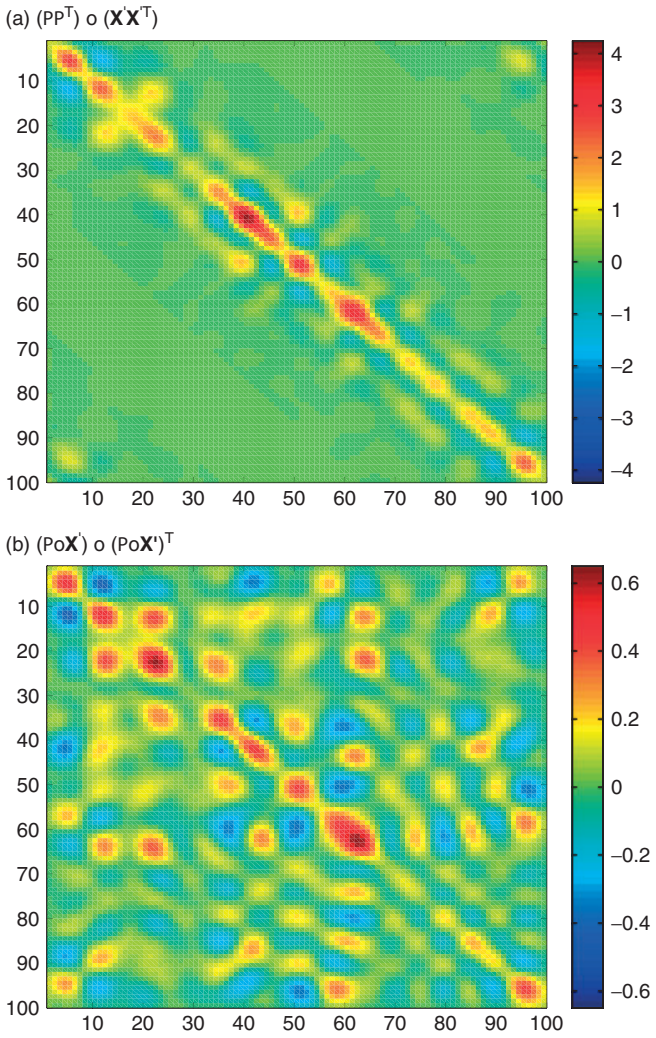

Figure 5. Covariance matrices produced from the model implemented in Petrie, 2008, using an ETKF to test the approximation of Schur product localization in a square root filter. ($\rho\,\rho^T$) ∘ ($\boldsymbol{X'}\,\boldsymbol{X'}^T$)

The process of data assimilation uses mathematical formulae to combine observational data with a model estimate, in accordance with their error distributions, to produce a more accurate analysis estimate; it is a key field of research in meteorology. Ensemble methods are becoming more popular as greater computing power becomes more affordable. Using ensembles that are too small, or undersampling, can cause additional problems such as underestimation of covariance, filter divergence and the development of long-range spurious correlations. Methods such as covariance inflation and covariance localisation have been developed in an attempt to overcome these problems; however, the implementation of some of these solutions is still an area of active research.

Acknowledgements

Ruth Petrie thanks the Royal Meteorological Society for providing an MSc scholarship which enabled this project to be completed at the University of Reading. Thanks also to the project supervisor Dr Sarah Dance for her support throughout. We would both like to extend our thanks to the reviewers for their very helpful comments, and to Dr Ross Bannister for all his help and advice.

References

Anderson JL, Anderson SL. 1999. A Monte Carlo implementation of the non-linear filtering problem to produce ensemble assimilations and forecasts. *Mon. Wea. Rev.* **126**: 2741–2758.

Anderson JL. 2001. An Ensemble Adjustment Kalman Filter for data assimilation. *Mon. Wea. Rev.* **129**: 2884–2903.

Bishop CH, Etherton BJ, Manjundar SJ. 2001. Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects. *Mon. Wea. Rev.* **129**: 420–436.

Bishop CH, Hodyss D. 2009. Ensemble covariances adaptively localized with ECO-RAP. Part 1: tests on simple error models. *Tellus* A **61**: 84–96.

Bjerknes V. 1911. *Dynamic meteorology and hydrography.* Part II. Kinematics. Carnegie Institute, Gibson Bros: New York.

Ehrendorfer M. 2007. A review of issues in ensemble-based Kalman filtering. *Meteorol. Z.* **16**: 795–818.

Evensen G. 2003. The Ensemble Kalman Filter: Theoretical formulation and practical implementation. *Ocean Dynamics.* **53**: 343–367.

Furrer R, Bengtsson T. 2007. Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *J. Multivariate Anal.* **98**: 227–255.

Hamill T, Whitaker JS, Synder C. 2001. Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Mon. Wea. Rev.* **129**: 2776–2790.

Houtekamer PL, Mitchell HL. 1998. Data assimilation using an Ensemble Kalman Filter Technique. *Mon. Wea. Rev.* **126**: 796–811.

Kalman R, Bucy K. 1961. New results in linear prediction filtering theory. Trans. AMSE J. *Basic Eng.* **83D:** 95–108.

Kalnay E. 2003. *Atmospheric modelling: Data assimilation and predictability.* Cambridge University Press: Cambridge.

Lorenc AC. 2003. The potential of the ensemble Kalman filter for NWP – a comparison with 4D-VAR. *Q. J. R. Meteorol. Soc.* **129**: 3183–3203.

Lorenz EN. 1963. Deterministic nonperiodic flow. *J. Atmos. Sci.* **20**: 130–141.

Oke P, Sakov RP, Corney SP. 2007. Impacts of localization in the EnKF and EnOI: experiments with a small model. *Ocean Dynamics.* **57**: 32–45.

Petrie R. 2008. *Localization in the Ensemble Transform Kalman Filter.* MSc Dissertation, University of Reading, Department of Meteorology.

Simmons A. 2003. *Observations, assimilation and the improvement of global weather prediction - Some results from operational forecasting and ERA-40.* ECMWF Seminar Proceedings: Recent developments in data assimilation for atmosphere and ocean.

Schur I. 1911. Bemerkungen zur theorie der beschrnkten bilinear formen mit unendlich vielen vernderlichen. *J. reine angew. Math.* **140**: 1–28.

Correspondence to: Ruth E. Petrie
Department of Meteorology, Earley Gate,
Whiteknights Campus,
University of Reading,
Reading, RE6 6BB, UK.

r.e.petrie@reading.ac.uk

DOI: 10.1002/wea.505