School of Mathematics, Meteorology
and Physics

University of Reading

Department of Mathematics

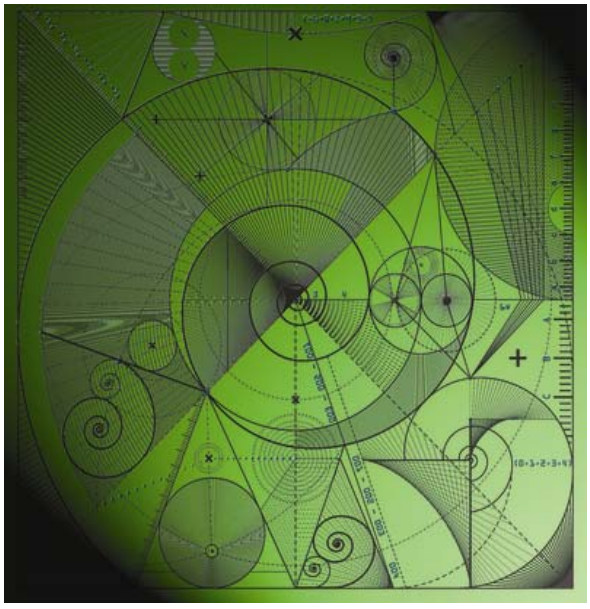
Preprint MPS_2010_19

21 April 2010

Ensemble data assimilation in the presence of cloud

by

S. Vetra-Carvalho, S. Migliorini, N.K. Nichols



Ensemble data assimilation in the presence of cloud

S. Vetra-Carvalho*, S. Migliorini, N. K. Nichols

School of Mathematics, Meteorology, and Physics
University of Reading, UK

Abstract

In numerical weather prediction (NWP) data assimilation (DA) methods are used to combine available observations with numerical model estimates. This is done by minimising measures of error between the observations and the model estimates, with more weight given to data that can be more trusted. For any DA method an estimate of the initial forecast error covariance matrix, $\mathbf{P}^f$, is required. For convective scale data assimilation, however, the properties of the error covariances are not well understood. An effective way to investigate covariance properties in the presence of convection is to use an ensemble based method for which the error covariance is readily available at each time step. In this work, we first investigate the performance of the ensemble square root filter (EnSRF) in the presence of cloud growth when applied to an idealised 1D convective column model of the atmosphere. We show that the EnSRF performs well to capture the cloud growth; however, the ensemble does not respond well to parametrized rain in the model. Secondly, we apply the EnSRF to the column model to investigate the properties of the ensemble error covariance matrix when convection is present.

Keywords:
Ensemble square root filter, convective data assimilation, covariance matrix, cloud fraction

1. Introduction

For operational meteorological centres, such as the Met Office, improving predictions of extreme weather is currently one of the main challenges. Such

*Corresponding author: s.vetra@reading.ac.uk

phenomena often impact on very localised regions and models are required to have high spatial resolution to maximise the chances of skillful forecasts. At high resolutions orography can be resolved much better, which allows the convection to be correctly organised in a geographical sense. Observations such as radar and high-resolution satellite measurements have also an essential role in improving model predictions when assimilated in the model. The aim of the data assimilation is to combine observations with model forecasts and to obtain the best possible estimate (also known as the analysis) of the atmospheric flow for the purpose of prediction.

A possible method for studying forecast errors is the Ensemble Square Root Filter (EnSRF), which is a statistical ensemble data assimilation method. In this method a set of state estimates (denoted as ensemble members) are used to represent an ensemble forecast error covariance. An ensemble can represent flow dependent uncertainties that vary in space and time. Hence, this approach can in principle provide better estimates than schemes with fixed covariance matrices. Various types of ensemble filtering schemes based on the Kalman Filter (KF) [3] have been proposed. Early ensemble methods were based on perturbed observations, in which each analysis ensemble member was derived from a Kalman-type equation using randomly perturbed observations. Using this approach the original Kalman Filter equations were satisfied statistically. However, perturbing observations adds an extra degree of uncertainty and, hence, several deterministic ensemble filters, including the EnSRF, were later derived. In these methods, it is required that the updated analysis perturbations satisfy the Kalman filter analysis error covariance equation. A very important benefit of the EnSRF (as for all ensemble methods) is that the error covariance matrix for these methods is readily available at each step with little extra cost, which is not the case for variational data assimilation techniques such as 3D-Var and 4D-Var. The ensemble preserves the nonlinearity of the model flow, as it does not linearise the dynamical model, and approximates the error covariance matrix of the state, thus becoming computationally efficient. Hence, the EnSRF not only provides the forecast at each time step, but also gives a probability of how good the forecast is.

All operational variational systems need to prescribe initial forecast error covariances and so far, for convective data assimilation systems, it is not known what properties this matrix should have. For example, the balance relations that are used to model forecast error covariances in variational data assimilation for synoptic scales are not necessarily suited for mesoscale flows

[9]. An ensemble could be used to investigate the relationships between model variables to estimate the initial covariance matrix for convective flow in variational systems. It is also important to study the performance of the ensemble method itself at convective scales, for example, the EnSRF performance in cases where the parameterized cloud growth is a strongly non-linear function of the state variables. Of particular interest, is to investigate the ensemble response to a regime switch from linear to highly-nonlinear (i.e. a switch from linear advection of state variables with no cloud to a sudden cloud growth in the system) and the EnSRF ability to trace the true solution in the presence of parametrized variables such as cloud fraction and rain.

Here we apply the EnSRF to an idealised 1+1D convective-scale model first developed by A. Rudd [5] with parametrized cloud and rain to investigate the performance of the EnSRF in the case of a sudden regime change from linear to highly non-linear. We examine the frequency of observations and the number of ensemble members needed for the EnSRF to capture the solution in the linear phase, as well as the ability of the ensemble to detect the switch and capture the cloud growth after the change of regime. We are especially interested in whether the cloud growth is indicated in the forecast error correlations before it actually happens. To test the ensemble further we consider the case where part of the growing cloud is allowed to rain out. This case is more complex as it correlates the two control variables, temperature and total water, which were previously independent of each other. It is also of interest to see if there is an optimal way to initialise the ensemble, which would give better results without increasing its size.

2. The EnSRF

An Ensemble Square Root Filter (EnSRF) has been built and implemented, as given by [2], making sure that the filter is unbiased and does not collapse [4], [1]. The EnSRF is based on the Kalman Filter (KF) equations,

$$\mathbf{x}^a = \mathbf{x}^f + \mathbf{K}\left(\mathbf{y} - \mathbf{H}\mathbf{x}^f\right) \quad (1)$$

$$\mathbf{P}^a = \left(\mathbf{I} - \mathbf{K}\mathbf{H}\right)\mathbf{P}^f \quad (2)$$

$$\mathbf{K} = \mathbf{P}^f\mathbf{H}^T\left(\mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{R}\right)^{-1} \quad (3)$$

where $\mathbf{x}$ is the state vector, $\mathbf{K}$ is the Kalman gain, $\mathbf{y}$ is the observation vector, $\mathbf{H}$ is the observation operator, $\mathbf{P}$ denotes an error correlation matrix and

superscripts a, f stand for analysis and forecast, respectively. The ensemble matrix is defined as

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N] \in \mathcal{R}^{n \times N} \tag{4}$$

where $\mathbf{x}_i$ are state vectors (ensemble members), n is the total number of variables in the control vector, and N is the number of ensemble members. The model state is assumed to be represented by the ensemble mean, which we define as

$$\overline{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}^{(i)} \qquad \text{or} \qquad \overline{\mathbf{X}} = (\overline{\mathbf{x}}, \overline{\mathbf{x}}, ..., \overline{\mathbf{x}})$$

and the ensemble perturbations are given by

$$\mathbf{X}' = \mathbf{X} - \overline{\mathbf{X}}.$$

The ensemble error forecast covariance matrix can be defined by

$$\mathbf{P}_e^f = \frac{1}{N-1} \mathbf{X}' \mathbf{X}'^T.$$

Note, that $\mathbf{P}_e \approx \mathbf{P}$, and as the ensemble number increases, $\mathbf{P}_e$ is expected to approach the true $\mathbf{P}$ as given by the KF. Thus, the update of the ensemble forecast error covariance (3) can be written as

$$(\mathbf{P}^a \approx) \frac{\mathbf{X}'^a \mathbf{X}'^{aT}}{N-1} = (\mathbf{I} - \mathbf{KH})\, \mathbf{P}_e^f.$$

Further, using the SVD decomposition of $\mathbf{KH}$, where $\mathbf{V}_2$ is the singular vector matrix and $\mathbf{\Sigma}_2$ is the singular value matrix, and the eigenvalue decomposition $\mathbf{HPH^T} + \mathbf{R} = \mathbf{Z\Lambda Z}^T$, where the columns of $\mathbf{Z}$ are eigenvectors and the diagonal elements of $\mathbf{\Lambda}$ are eigenvalues, we can express the analysis ensemble perturbations explicitly as

$$\mathbf{X}'^a = \mathbf{X}'^f \mathbf{V}_2 \sqrt{\mathbf{I} - \mathbf{\Sigma}_2^T \mathbf{\Sigma}_2} \mathbf{V}_2^T. \tag{5}$$

Note that since $\mathbf{V}_2^T$ is a symmetric orthogonal matrix, then $\mathbf{V}_2^T \mathbf{V}_2 = \mathbf{I}$ and post-multiplying equation (5) by $\mathbf{V}_2^T$ does not alter the analysis error covariance, $\mathbf{P}_e^a$. However, this rotation keeps the filter unbiased [6]. This gives us the analysis for the state (the ensemble mean)

$$\overline{\mathbf{x}}^a = \overline{\mathbf{x}}^f + \mathbf{X}'^f \mathbf{S}^T \mathbf{Z\Lambda}^{-1} \mathbf{Z}^T \left(\mathbf{y} - H\left(\overline{\mathbf{x}}^f \right) \right), \tag{6}$$

where the terms $\mathbf{V}_2, \mathbf{\Sigma}_2, \mathbf{S}, \mathbf{Z}, \mathbf{\Lambda}$ come from the decompositions (see [2] for more detail). Finally the analysis of the ensemble is

$$\mathbf{X}^a = \overline{\mathbf{X}}^a + \mathbf{X}'^a \tag{7}$$

and the ensemble analysis forecast error covariance matrix is

$$\mathbf{P}_e^a = \frac{1}{N-1}\mathbf{X}'^a\mathbf{X}'^{aT}. \tag{8}$$

The main difference between the traditional EnKF and the square root version EnSRF (and the reason for using EnSRF) is that in the square root algorithm the perturbation of measurements is avoided as this can lead to more errors. Also, we do not need to perform inversion of the observation error correlation matrix, $\mathbf{R}$, nor do we need the assumption of uncorrelated measurement error covariance matrix [1] [2].

3. 1+1D column model

In this work we use a 1+1D model (1D in space and time) describing the atmospheric flow in a vertical column. The model state variables are vertical velocity, $w(z)$, temperature, $T(z)$, total water, $q_t(z)$, pressure, $p(z)$, temperature change with height, $\Psi(z)$, liquid cloud water, $q_{cl}(z)$, saturated vapour, $q_{sat}(z)$, and cloud fraction, $f(z)$, where $z \in [0, 12]$ km is vertical height. From all of the model variables, only $T(z)$ and $q_t(z)$ are used in the data assimilation process to update the system, and these are known as the control variables. In vector form we define the control vector $\mathbf{x}$, as

$$\mathbf{x} = \begin{pmatrix} \mathbf{T} \\ \mathbf{q}_t \end{pmatrix} \in \mathcal{R}^{102\times 1}, \tag{9}$$

with height z being discretised with 51 equal levels and $\mathbf{T},\ \mathbf{q}_t\ \in\ \mathcal{R}^{51\ \times\ 1}$.

Control variables, T and q_t, at a given height z are linearly advected by a known vertical velocity $w(z) = 0.5 \sin((z/z_{top})\pi)$, constant in time with maximum speed in the middle atmosphere of 0.5 m/s. The model uses a cloud scheme [7], to compute a strongly non-linear cloud fraction, f, given by

$$f(z) = 0.5\left(1 + \tanh\left(\frac{2q_{cl}(z)}{q_{sat(z)}(1 - RHc)}\right)\right), \tag{10}$$

[1] Here we do use uncorrelated $\mathbf{R}$.

where RHc is critical relative humidity, $q_{cl} = q_t - q_{sat}$ and $q_{sat} = \epsilon e_s/p$ with $\epsilon = 0.622$ being the ratio of molecular weights of water and dry air and $e_s = e_s(T)$ being the saturation vapour pressure. Thus, cloud fraction, f, depends on both control variables T and q_t, and the range of f is [0,1] with $f = 0$ meaning no cloud and $f = 1$ meaning full cloud.

The model exhibits linear and non-linear regimes. If the model exhibits no cloud, the system is in a linear state and we refer to this as a 'no cloud regime' or 'linear regime'. However, if the model has cloud growth, the system is in a non-linear state and we refer to this as a 'cloud regime' or 'non-linear regime'.

3.1. Rain parametrization

The threshold for cloud fraction at which the rain parametrization is switched on is $f > 0.2$. When this threshold is reached we reduce the cloud fraction by using a parameter $r \in (0, 1)$ as follows,

$$f_r = rf. \tag{11}$$

The rain fraction parameter varies smoothly to reduce the cloud fraction over an hour. [2] Next we invert equation (10) to express it in terms of liquid cloud water as

$$q_{cl} = 0.5 q_{sat} \left(1 - RHc\right) \operatorname{arctanh}\left(2f\right). \tag{12}$$

Hence, using the reduced cloud fraction (11) in equation (12) we obtain a new liquid cloud water value $q_{cl,r} = q_{cl}(f_r)$. The vapour amount, q_{sat}, is computed from the temperature profile as in the case of no rain, and the reduced total water amount is given by

$$q_{t,r} = q_{cl,r} + q_{sat} \tag{13}$$

for each vertical level where $f > 0.2$ in each ensemble member.

4. Results of experiments

Using our 1+1D model, we aim to investigate the performance of the EnSRF in the case of a sudden regime change from near-linear to highly

[2] Note that, depending on the cloud growth, the cloud fraction f can increase faster than the rain parameter r can reduce it.

non-linear. The model parameters are chosen so that initially for the first 3 hours there is no cloud in the solution, but then a very sudden and fast cloud growth occurs in the upper half of the atmosphere. To examine the EnSRF performance in these conditions we use a twin experiment. We create a 12 hour long reference solution ('truth') from initial chosen profiles of $T(z, 0)$ and $q_t(z, 0)$. We then sample observations with some time frequency from the reference solution and add noise with a $\sigma_o = 1\%$ variance. We also perturb the initial profiles with $\sigma_e = 10\%$ variance to give an initial ensemble mean around which we create the initial ensemble. Using a twin experiment allows us to deduce the accuracy of the ensemble estimate as a difference between the ensemble analysis and the truth.

4.1. Results for regime switch

The ensemble exhibits a high ability to capture the solution in the linear phase (i.e. no cloud) even with a small ensemble size ($N = 10$) with respect to the size of the state space ($n = 102$), provided enough good observations of both control variables are given at a suitable time frequency, e.g every 30 min. For larger ensemble size the observation quality and/or frequency can vary to obtain the same accuracy of the solution. Interestingly, the EnSRF is able to capture the regime switch in most cases, with the accuracy and the rate of ensemble convergence to the true solution depending mainly on the ensemble size and secondly on the observation frequency. However, it is important that both control variables are observed at least once before the cloud growth. Note that, in cases where the ensemble size is small, we find that the EnSRF over-predicts the cloud growth, i.e. in the ensemble estimate, cloud growth is initiated sooner than in the reference solution. This clearly is an issue, especially in an operational setting, where the ensemble size is much smaller than the size of the state space. However, in this idealised model, unless liquid is removed from the system, the model would just saturate in the levels where cloud is present, thus becoming linear again. If this is allowed, the ensemble, even with $N = 10$, can capture the solution very well after a few observation cycles.

4.2. Results with rain parametrization

To keep nonlinearity in the system we introduce parametrized rain as explained in section 3. This reduces the cloud fraction after the cloud has grown to a given threshold in the column. In this case, a large ensemble (say $N > 102$) will be able to capture the solution with parametrized rain

well if good observations (at least every 30min) are used. However, it is not realistic, for practical applications, to have an ensemble size larger than the state space. With a small ensemble size and many good observations at the initial phase (no cloud regime), the ensemble is not capable of increasing its spread to respond when the parametrized rain is switched on/off. This can be seen in figure 1, where the ensemble spread is larger than the ensemble analysis error until the parametrized rain is switched on. This indicates that the ensemble spread captures the true solution and thus has the ability to trace it. We also see that for the cloud fraction, f, the ensemble has 'learned' after the first set of observations that there is no cloud in the system in the first 3 hours, i.e. both the analysis error (blue) and ensemble spread (red) for f become very small for this period. The combination of a small ensemble size and many good observations assimilated in the 'no-cloud' regime results in the ensemble spread becoming too small to respond when the parametrized rain is switched on. We see from figure 1 that a small ensemble, here $N = 30$, is not able to increase its spread to capture the true solution when the rain parametrization is switched on. The ensemble spread keeps on decreasing (red), whereas the analysis error increases (blue). On the other hand, if only a limited portion of the state space is observed, say every tenth vertical level for both control variables, then the small ensemble has a spread which encompasses the true solution even after the parameterized rain is switched on. This can be seen from figure 2, where the ensemble spread responds (in red) to the parametrized variables and the ensemble analysis error (in blue) is always below the ensemble standard deviation. We note that, even though the ensemble has 'learned' that there is no cloud in the first 3 hours, the accuracy in this period is much worse than that of the ensemble that used more good observations, as can be seen from the plots for cloud fraction, f, in figures 1 and 2.

4.3. Forecast error correlations

We are also very interested in whether the cloud growth is indicated in the forecast error correlations before it actually happens. First of all we study the theoretical correlations between T and q_t by looking at the model. It is possible to see that in the model there are no cross-correlations between T and q_t if neither cloud nor rain is present in the system. However, variables are cross-correlated in the model when cloud develops and/or rain is permitted.

In the ensemble correlation matrix, even if we start with a small ensemble that has spurious cross-correlations between the control variables, these

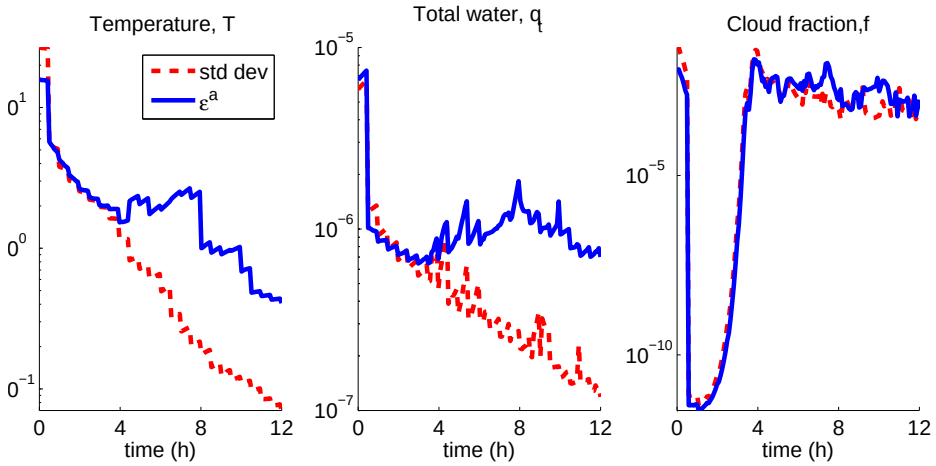


Figure 1: Ensemble analysis error (blue) compared to ensemble standard deviation (red). Observing all variables every 30min, $N = 30$. Rain at 3h.

are eliminated after a few assimilation cycles (how many assimilation cycles it would take depends on the ensemble size) so that the control variables become uncorrelated. Small cross-correlations appear on the levels where cloud develops and, since for small ensemble sizes cloud develops sooner in the ensemble estimate than in the reference solution, the ensemble correlation matrix is not accurate until cloud also develops in the reference solution. When the rain is permitted in ensemble members, the cross-correlations of T and q_t are very strong at the locations where it is raining and clearly if the ensemble is too small to increase its spread then its cross correlations will not be accurate.

5. Conclusions

To investigate the ensemble performance in a convective system, we have applied the EnSRF to a 1+1D column model with parametrized cloud and rain. We have examined the ability of the ensemble to capture a sudden regime switch. In our idealised model this was represented by a near-linear (no cloud) regime for the first three assimilation hours followed by a sudden cloud growth (non-linear regime). This was intended to represent a sudden development of convection in the system.

We showed that the EnSRF, even with a small number of members, was able to capture the true solution in the linear phase. Moreover, as long as the

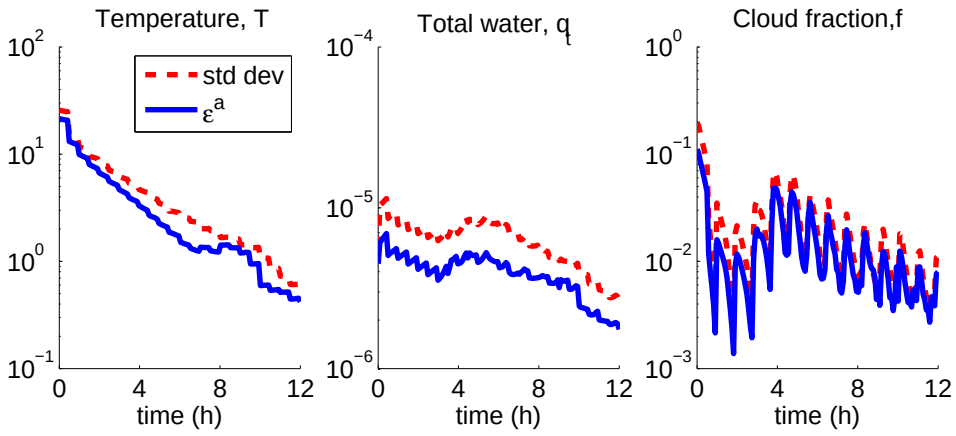


Figure 2: Ensemble analysis error (blue) compared to ensemble standard deviation (red). Observing every tenth level of T and q_t every 30min, $N = 30$. Rain at 3h.

system was observed at least once before the cloud developed, even with a small ensemble size, the EnSRF was able to detect the regime switch. However, the accuracy with which the switch was captured depended mainly on the ensemble size and secondly on the observation frequency. An important note is that for small ensemble sizes the EnSRF developed cloud in its solution sooner than in the reference solution. The reason for this and the impact it has on the ensemble solution needs to be investigated further.

In the cases where parametrized rain was permitted after the initial cloud development, we found that for ensembles that have more ensemble members than variables with good frequent observations, the performance of the EnSRF was good. By this we mean that the ensemble spread was always larger than the ensemble error, thus capturing the true solution within its spread. However, as in operational settings, if the ensemble size was much smaller than the state space and many good observations were used, the ensemble became too confident in the linear regime and was unable to respond when the rain was switched on. This was rectified by having fewer observations, i.e. only observing every tenth level of the two control variables. By doing this, the ensemble spread was always larger than its error. However, the accuracy of the ensemble solution in the linear phase was obviously worse than with many observations.

Finally, we have performed an initial investigation of the ensemble correlation matrix in convective conditions. The ensemble develops correlations and cross-correlations as expected between the variables. However, with small en-

sembles and, in particular, when rain was parametrized, the correlations and cross-correlations at the time of cloud development were shifted in time or incorrect as the ensemble either developed cloud too soon and hence developed cross-correlations too soon, or its spread was too small, thus resulting in incorrect correlations.

References

[1] L. Baker, *Properties of the Ensemble Kalman Filter*, 2007, University of Reading, Department of Mathematics, MSc Thesis

[2] G. Evensen, *Sampling strategies and square root analysis schemes for the EnKF*, Ocean Dynamics, 2004, Vol 54, pp 539 - 560

[3] R.E. Kalman, *A new approach to linear filtering and prediction problems*, 1960, Journal of Basic Engineering, Vol. 82 (1), pp. 3545

[4] D. Livings, S. L. Dance, N., K., Nichols *Unbiased ensemble square root filters*, 2008, Physica D, Vol 237, pp. 1021 - 1028

[5] A. C. Rudd, *The effect of nonlinearity on the variational assimilation of satellite observations using a simple column model*, 2009, PhD Thesis, University of Reading

[6] P. Sakov, P. R. Oke, *Implications of the form of the ensemble transformation in the ensemble square root filters*, 2007, Mon. Weath. Rev., Vol. 136, pp. 1042-1053

[7] R. N. B. Smith, *A scheme for predicting layer clouds and their water contents in a general circulation model*, 1990, Q. J. R. Meteorol. Soc., Vol 116, pp. 435 460

[8] M. K. Tippett, J. L. Anderson, C. H. Bishop, T. M. Hamill, J. S. Whitaker, *Ensemble Square Root Filters*, 2003, AMS, Mon. Weath. rev., Vol. 131, No 17, pp 1485 - 1490

[9] S. Vetra-Carvalho, M. Dixon, S. Migliorini, N.K. Nichols, S. Ballard, *Breakdown of hydrostatic balance at convective scales in the forecast errors in the Met Office Unified Model*, 2010, University of Reading, Department of Mathematics, Preprint Series, MPS-2010-10