# A partial ensemble Kalman filtering approach to enable use of range limited observations

**Borup, Morten; Grum, Morten; Madsen, Henrik; Mikkelsen, Peter Steen**

*Published in:*
Stochastic Environmental Research and Risk Assessment

*Link to article, DOI:*
10.1007/s00477-014-0908-1

*Publication date:*
2015

*Document Version*
Publisher's PDF, also known as Version of record

*Citation (APA):*
Borup, M., Grum, M., Madsen, H., & Mikkelsen, P. S. (2015). A partial ensemble Kalman filtering approach to enable use of range limited observations. Stochastic Environmental Research and Risk Assessment, 29(1), 119-129. DOI: 10.1007/s00477-014-0908-1

**Corrected Proof.**

Final version published in *Stochastic Environmental Research and Risk Assessment* Vol. 29, Issue 1, pp 119-129 (http://link.springer.com/article/10.1007%2Fs00477-014-0908-1)

# A Partial Ensemble Kalman Filtering approach to enable use of range limited observations

Morten Borup[1+2*], Morten Grum[2], Henrik Madsen[3], Peter Steen Mikkelsen[1].

[1] Department of environmental Engineering. Technical University of Denmark
[2] Krüger A/S, Veolia Water Solutions and Technologies, Denmark
[3] DHI, Denmark

* Corresponding author: morb@env.dtu.dk, Telephone: +45 45422182.

## Abstract

The ensemble Kalman filter relies on the assumption that an observed quantity can be regarded as a stochastic variable that is Gaussian distributed with mean and variance that equals the measurement and the measurement noise, respectively. When a gauge has a minimum and/or maximum detection limit and the observed quantity is outside this range, the signal from the gauge can, however, not be related to the observed quantity in this way. The current study proposes a method for utilizing this kind of out-of-range observations with the ensemble Kalman filter by explicitly treating the out-of-range observations. By doing this it is possible to update the ensemble members that are within the observable range of the gauge towards the observation limit and thereby reduce the ensemble spread. The method is tested using both a linear and a non-linear simple forcing-driven model in perfect model experiments where the same model and noise descriptions are used for the truth simulation and for the ensemble Kalman filter. The results show that the positive impact of the method in case of range-limited observations can exceed that of increasing the ensemble size from 10 to 100 and that the method makes it possible to improve model forecasts using observations that would otherwise have been non-informative.

**Keywords:** Data assimilation, ensemble Kalman filter, observation limit, range-limited observations.

## 1 Introduction

Most environmental modelling is associated with substantial uncertainties caused by model errors as well as uncertainties in input data. These uncertainties can be reduced by means of data assimilation (DA) that adjusts the model using measurements that are related to some of the modelled quantities. One of the most popular data assimilation methods is the Ensemble Kalman Filter (EnKF) introduced by Evensen (1994) as a flexible and efficient alternative to the Extended Kalman Filter for large non-linear models. The method can be seen as a Monte Carlo implementation of the classical Kalman filter  (Kalman, 1960) in which an ensemble of models is used to represent the error statistics. By doing this all the non-linearities in the model are included in the forward propagation of the error and are thereby implicitly included in the model update. The model update is linear and is based on the ensemble covariance calculated around the ensemble mean, which means that the filter is only optimal for strictly linear systems with Gaussian errors. Nonetheless, the EnKF has proven very successful for non-linear models within a wide range of applications such as oceanography, meteorology, oil reservoir modelling, groundwater modelling, hydrology, etc. (Keppenne and Rienecker, 2002; Lee et al., 2012; Nævdal et al., 2003; Olume, 2006; Tong et al., 2012) and must be regarded as one of the most versatile DA methods available.

In data assimilation the model is combined with information from multiple data sources. Every new independent observation can be used to improve the model accuracy and reduce the model uncertainty. This means that it is a big advantage if a DA scheme is capable of utilizing as many different kinds of observations as possible. Many of the measurements available in environmental systems are only defined within a certain interval. Examples of downwards limited measurements are any kind of concentration measurements, river water level measurements from satellite radar altimetry, float or pressure water level measurements, oil or ground water reservoir levels from borehole data, etc. Some measurements are furthermore only available within a limited interval of the actual variation of the quantity. An example of this is satellite estimated thickness of the ocean ice cover that can be used in global climate models (Kaleschke et al., 2010). The satellite data can be used to determine whether there is an ice cover or not but can only quantify the ice thickness up to approximately half a meter for the Arctic where the actual ice thickness can grow to several meters. Some measurements can even be of Boolean nature such as information about water or oil wells being empty or not,  or if there is overflow or not at a weir in urban hydrology (Thorndahl et al., 2008). Optimally all these data sources provide information about the state of the physical systems and therefore have the potential to improve model predictions. Most data assimilation methods, however, do not work in the absence of actual quantifiable observations. It has not been possible to find references in the data assimilation literature on the use of range-limited observations, which suggest that practice is either to not use data from gauges that are not continuously covering the observed quantities or to only perform the assimilation in the periods where quantifiable observations are available. This is, however, a waste of valuable information. When a gauge does not return a signal that is within its observable range, it provides the valuable information that the observed quantity is probably not within this range. An optimal data assimilation scheme should preferably be capable of using this information.

The current study describes a new method for utilizing the information available in out-of-range observations when using the EnKF. The key element of the method is, in the absence of observations within range, to define a virtual observation at the limit of the observation interval. This artificial observation is then used to correct only the ensemble members that are within the observation interval even though the lack of observations suggests that the observed quantity is outside. This process is referred to as *partial updating* since usually only a part of the ensemble is updated. The justification and description of the method is described in the section "Partially updating ensemble", where it is shown that the method in a consistent way enables the EnKF to operate with non-Gaussian data likelihoods of out-of-range observations. In the "Numerical Tests" and "Results and Discussion" sections the effect of the method is tested using both linear and non-linear reservoir cascade models, which could represent many forcing-driven environmental systems. The method should, however, be valid for all types of models.

# 2 Background

## 2.1 The Ensemble Kalman Filter

In the following an overview of the most import parts of the EnKF will be given. A more thorough description can be found in (Evensen, 2003).

The Kalman filter can be seen as a subset of Recursive Bayesian Estimation for linear models in which all variables are assumed Gaussian distributed, which means that the Kalman filter only works with the mean and (co)variance. The EnKF uses an ensemble of state-space models to represent the background error covariance that is needed for the Kalman filter. The ensemble is created by perturbing model forcing, model parameters and/or model states based on the modeller's assumptions/knowledge of their errors. If an ensemble has *n* members, it can be written as:

$$\boldsymbol{X} = [\boldsymbol{x}_1, \dots\dots, \boldsymbol{x}_n] \tag{1}$$

where $\boldsymbol{x}_i$ is the full state vector of the *i*'th ensemble member.

Every time a new observation is available the ensemble is updated using this observation. The update state is called the analysed state and is in the following denoted with the superscript *a*. The model is initialised with the updated ensemble and all the ensemble members are propagated forward in time until the next analysis. The model forecast is called the background state and will in the following be denoted with a superscript *b*. The EnKF analysis consists of applying the Kalman filter analysis equation separately to each member of the ensemble:

$$\boldsymbol{x}_i^a = \boldsymbol{x}_i^b + \boldsymbol{K}\left(\boldsymbol{d}_i - \boldsymbol{H}\boldsymbol{x}_i^b\right) \tag{2}$$

where $\boldsymbol{K}$ is the Kalman gain, $\boldsymbol{H}$ is the measurement operator that maps the observations to the state variables, and $\boldsymbol{d}_i$ is the vector of observations used to update the *i*'th ensemble member created by perturbing the actual observation vector **d**.

***K*** can be calculated from the background error covariance $\boldsymbol{P}^b$ and the observation error covariance ***R***:

$$\boldsymbol{K} = \boldsymbol{P}^b\boldsymbol{H}^T\left(\boldsymbol{H}\boldsymbol{P}^b\boldsymbol{H}^T + \boldsymbol{R}\right)^{-1} \tag{3}$$

The main idea behind the EnKF is to replace the error covariance with the ensemble covariance which can be calculated from the ensemble. This is, however, not necessary to do explicitly since the gain can be computed much more efficiently directly from the ensemble without actually calculating and storing the ensemble covariance. Ways to do this are described in (Houtekamer and Mitchell, 2001; Sakov et al., 2010).

When an update is performed, the ensemble spread is reduced and the ensemble mean moves towards the observation as illustrated in Figure 1 (Left). This process is repeated every time a new observation is available, see Figure 1 (right).

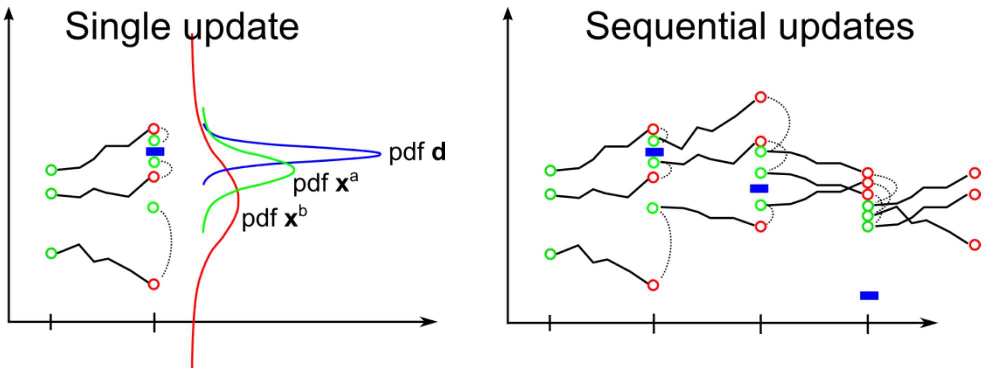

Figure 1: Single (left) and sequential (right) updates. The red dots are the background state values while the green dots are the state values after the updates. The rectangular blue marks are the observations.

For Gaussian distributed variables the maximum-likelihood estimate equals the minimum-variance estimate (van Leeuwen and Evensen, 1996), which means that by regarding the ensemble covariance as being an efficient estimate of the background error covariance for each of the individual ensemble members, these can efficiently be updated using the Kalman filter. The ensemble of updated members represents the posterior distribution. If all the ensemble members are updated towards the same observed value using the standard Kalman update equation (2), then the ensemble spread is reduced too much compared to the theoretical optimal value. In order to overcome this problem the standard EnKF updates each ensemble member with independently perturbed observations with mean equal to the observed value and a variance that reflects the observation uncertainty  (Burgers et al., 1998; Houtekamer and Mitchell, 1998).

## 2.2 Deterministic EnKF

All ensemble based DA methods are affected by sampling errors due to the use of ensembles to represent probability distributions. The fact that the standard EnKF relies on perturbed observations results in additional sampling errors – especially in case of small ensemble sizes. This problem will be even more pronounced when updating only a part of the ensemble, as suggested in the current work. To address the sampling error problem Sarkov and Oke (2008) presented the *Deterministic EnKF* (DEnKF) which is a deterministic formulation of the EnKF in the sense that the observations are not perturbed. Instead the update is divided into two steps where, firstly, the ensemble mean is updated separately as in the regular EnKF (see above) and, secondly, the ensemble anomalies are updated using only half the gain in order to avoid excessive reduction in the ensemble spread. Proof of this approach can be found in Sarkov and Oke (2008).

The anomalies are the deviations from the ensemble mean calculated as

$$\boldsymbol{A}_i = \boldsymbol{x}_i - \overline{\mathbf{X}} \tag{4}$$

where $\overline{\mathbf{X}}$ is the ensemble mean state vector. Once the anomalies are calculated the mean is updated using equation (2) and the anomalies are updated using:

$$\boldsymbol{A}^{a} = \boldsymbol{A}^{b} - \tfrac{1}{2}\boldsymbol{KHA}^{b} \qquad (5)$$

Hereafter the updated ensemble can be reconstructed by adding the updated anomalies to the updated mean state vector. The DEnKF is comparable to the ensemble square root filter (EnSRF) implementation of Whitaker & Hamill (2002), which uses a very similar analysis scheme, but calculates a factor $\alpha$ that is used instead of the factor of ½ in equation (5), in order to obtain a solution that exactly matches the theoretical optimal analysed error covariance in the linear case. The similarities are especially clear in the case of a single observation, in which case α is a scalar that converges from 1 towards ½ as the ensemble background error variance for the observed location decreases compared to the variance of the observation error. This implies that the DEnKF overestimates the analysed error covariance, with the largest overestimation when the observation error is relatively small, which can be seen as an implicit inflation of the ensemble.

# 3 Partially Updating Ensemble

In case of an out-of-range observation (OR-observation) it is comprehensible that the ensemble members that are in fact outside the observable range should not be updated. What would be a reasonable update for the members that are inside the observable range is a much less intuitive question to answer. This will be explored in the following section.

## 3.1 Posterior maximum likelihood estimate of individual ensemble members given out-of-range observations

The EnKF can be seen as a Markov Chain Monte Carlo implementation in which the members of the ensemble, that describes the probability density of the true state, are conditioned on new data using recursive Bayesian estimation.  The conditioning is performed by the Kalman filter, which results in the maximum likelihood estimate when assuming Gaussian errors and data likelihood.

The Kalman gain that is used to update each of the ensemble members is calculated from the ensemble statistics. This means that each ensemble member implicitly is attributed the ensemble variance as *a priori* variance, which is weighted against the observation variance in the process of estimating the posterior mean, as an approximation of the maximum likelihood estimate. By looking at the ensemble update in this way - as the process of finding the maximum of an implicit posterior distribution - it becomes possible to determine a reasonable and consistent update in the case of out-of-range observations by making a few assumptions about the OR-observation likelihood.

First of all it is assumed that the likelihood is constant outside the observable range, since an OR-observation in itself does not contain any information about the specific value of the observed quantity. Secondly, it is assumed that the shape of the likelihood function from the detection limit and into the observable range is determined by the observation uncertainty. This corresponds to assuming that there is a chance of the quantity not being observed even though it is within the observable range, and that this chance increases the closer the quantity is to the detection limit, according to the uncertainty of the gauge. An illustration of such an OR-observation likelihood function, in the case of a lower observation limit, can be seen in Figure 2.

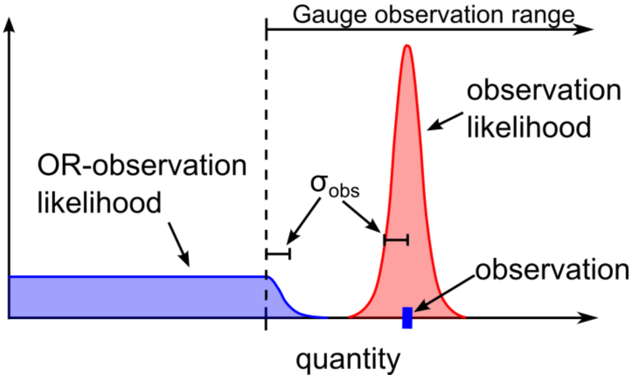

Figure 2: Data likelihoods when the gauge has a lower observation limit given an actual observation (red) and given an OR-observation (blue). $\sigma_{obs}$ is the standard deviation of the observation.

According to Bayes rule the posterior distribution is proportional to the prior distribution multiplied with the likelihood function of the new data. In the case of an actual observation this is a product of two Gaussian distributions and is therefore itself Gaussian. This is not the case when using the OR-observation likelihood function that is illustrated in Figure 2. Figure 3 shows the implicit prior and posterior distributions of two members, one outside and one inside the observable range, being updated by the OR-likelihood. The location of the posterior maximum for member 2 will not change since the prior maximum coincides with the maximum of the likelihood function. This means that members outside the observable range should not be affected by the update given an OR-observation. Note that the mean of the posterior would be a poor estimator, since this would imply that the member would be adjusted away from the detection limit continuously whenever no observations are present.

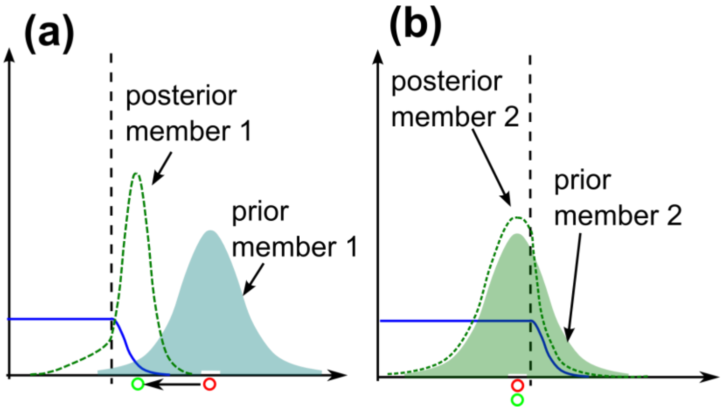

Figure 3: Implicit posterior distributions (dotted lines) of individual members conditioned on the OR-observation likelihood and Gaussian priors in the case where the member is inside (a) and outside (b) the observable range, respectively. The solid blue curve is the OR-observation likelihood when using a gauge with a lower detection limit.

The situation for the member that is inside the observable range (Figure 3a) is quite different, but the location of the posterior maximum can still be deduced in a simple manner. Since the detection limit is located at the likelihood maximum, the mode of the posterior will be located somewhere between the detection limit and the mode of the prior. This means that the location of the posterior maximum solely is determined by the shape of the prior and the part of the likelihood function that is inside the observable range. Following the assumption that the shape of the likelihood function within the observable range is described by the observation uncertainty used for the actual observations, the posterior mode will be located the same place as if an actual observation was at the detection limit. This is very convenient since it means that the members that are within the observation interval, in case of an OR-observation, can be updated using the same analysis equations as when an actual observation is present, by simply assuming an observation at the detection limit. Meanwhile, the members that are outside the observable range should be left untouched.
)

To sum up, this means that when assuming an OR-observation likelihood as the one shown in Figure 2 and assuming Gaussian priors for all ensemble members, as it is implicitly done in the EnKF, the most likely value of the ensemble members can be found by only updating the ensemble members that are inside the observable range towards a virtual observation at the detection limit. The method is consistent in the sense that all ensemble members are conditioned on the same likelihood function. Note that the integral of the OR-observation likelihood function can be infinite, but since this integral is not computed in Bayes theorem it does not affect the general applicability of the method. Besides, the same posterior maximum will be found if the OR-observation likelihood function is set to zero somewhere outside the span of the ensemble, cf. Figure 3, which means that a finite integral can be obtained by setting the likelihood to zero for values above a threshold that is far from what will ever be observed for the quantity in question.

## 3.2 Implementation

For a well-functioning EnKF setup the forecasted background ensemble should span over the true state for the vast majority of the time. This implies that in most cases only part of the ensemble can be expected to be within the observable range in case of an OR-observation. Therefore only a part of the ensemble should be updated according to the procedure described above, which means that the update process would be very prone to suffer from sampling errors if based on the standard EnKF formulation using perturbed observations. Therefore the DEnKF formulation of the EnKF has been chosen as basis for the partial updating.

The implementation is made by assuming a virtual observation at the detection limit whenever the observations are out of range. Since the members that are outside of the observable range should be left untouched, the mean should not be changed explicitly and only parts of the anomalies should be corrected. The updating scheme is conditioned upon the in- or out-of-range status of the observations, such that the ordinary DEnKF updating scheme is used in case of an actual observation while the partial updating scheme is used otherwise.

The main part of the implementation is to construct the $n_{obs}$ x $n_{ens}$ innovation matrix ***C*** for the updating of anomalies, where $n_{obs}$ and $n_{ens}$ are the number of observations and number of ensemble members, respectively. The value of an element in ***C*** for a given observation and ensemble member determines, together with the Kalman gain, the change to the given ensemble member. In case of actual observations the DEnKF scheme is used which means that ***C*** contains the ensemble members' departure from the ensemble mean at the observed location. In case of OR-observations the values in ***C*** equals the individual ensemble members' departure from the observation limit, as long as these members are within the observable range – otherwise the values are set to zero. The work flow for this is described with pseudo code and equations below. For simplification the example is for a single observation point only.

### Analysis start

```
if $d$ is an actual observation
        $\boldsymbol{C} = \boldsymbol{H}\boldsymbol{A}^{\boldsymbol{b}}$                (6)
else
    for each member $i$
```

```
          if $\boldsymbol{Hx}_i$ is within observable range
          $\boldsymbol{C}_i = \boldsymbol{Hx}_i - limit$          (7)
          else
          $\boldsymbol{C}_i = 0$          (8)
          end if
      end for each
  end if-else
          $\boldsymbol{A}^a = \boldsymbol{A}^b - \frac{1}{2}\boldsymbol{KC}$          (9)
  when $d$ is within range only:
          $\overline{\boldsymbol{X}}^a = \overline{\boldsymbol{X}}^b + \boldsymbol{K}(\boldsymbol{d} - \boldsymbol{H}\overline{\boldsymbol{X}}^b)$          (10)
Analysis end
```

Notice that in the case of actual observations the equations used are 6, 9 and 10, which correspond to the standard DEnKF updating scheme. A multiple observations implementation has to acknowledge that not all observations are within range at the same time. This can be done by updating the mean every time, but setting the innovations for the mean to zero for OR-observations.

When sequentially applying the partial update to an ensemble, the distribution is likely to become skewed since only a part of the ensemble is updated and this is always in the same direction, see illustration in Figure 4. This means that the spread of the ensemble is reduced but at the cost of violating the Gaussian assumption behind the EnKF. The purpose of the EnKF is, however, not to produce Gaussian error estimates, but to produce the best error estimates with the information available, and the EnKF has proven to be efficient even for non-linear systems with non-Gaussian errors. The fact that the ensemble spread is reduced closer to the true value of the state makes it possible for the EnKF to estimate the covariance closer to the true state, which should provide better updates once an actual measurement becomes available.

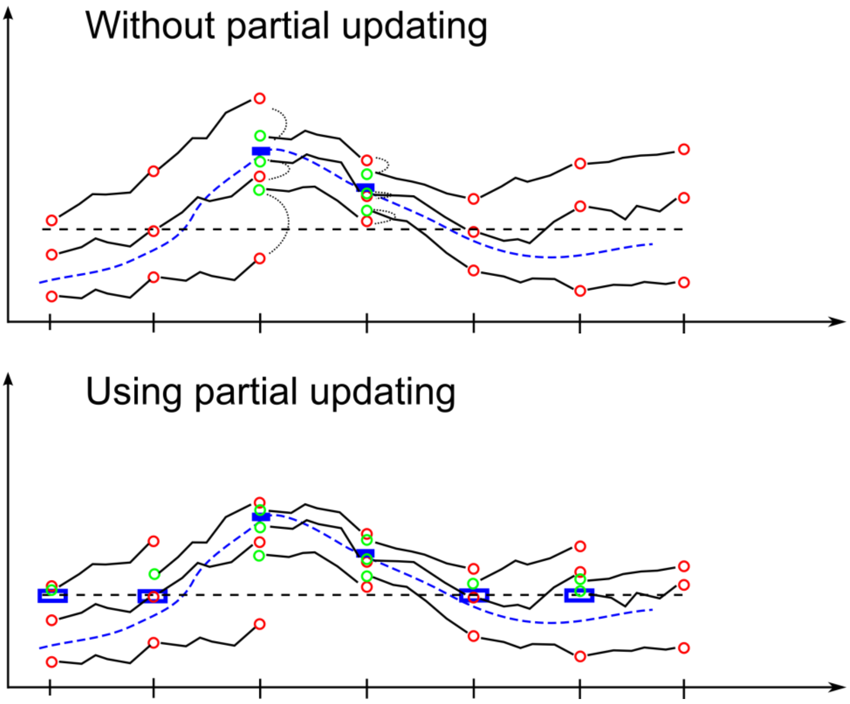

**Figure 4: Sequential ensemble updating with and without the use of partial updating when a gauge has a lower observation limit at the horizontal black dashed line. The solid blue rectangular marks are actual measurements while the empty blue rectangular marks indicate the virtual measurements used for the partial updating when no actual measurements are available. The red dots are the background state values while the green dots are the state values after the updates. The dashed blue lines show the true solution and the thin dotted black lines in the top show which posterior belongs to which prior.**

# 4 Numerical Tests

## 4.1 Models

In this section the proposed partial updating scheme is illustrated on two simple reservoir cascade models. The models are inspired from the simple forecast models often used in hydrology (Aubert et al., 2003; Birkel et al., 2010; Löwe et al., 2013; Thorndahl et al., 2013) but are likely to resemble many forcing-driven environmental models. Both models consist of three linear or piecewise linear reservoirs with forcing applied to the first reservoir only while the observations are on the last reservoir. The dimensionless forcing for both models follows a Gamma distribution with shape parameter of 0.01 and a scale parameter of 100. The models are integrated forward in time using the standard fourth-order Runge-Kutta method.

### 4.1.1 Linear model

The linear model is defined as:

$$d\begin{bmatrix}\boldsymbol{x}_1\\ \boldsymbol{x}_2\\ \boldsymbol{x}_3\end{bmatrix}=\begin{bmatrix}F-k\boldsymbol{x}_1\\ k\boldsymbol{x}_1-k\boldsymbol{x}_2\\ k\boldsymbol{x}_2-k\boldsymbol{x}_3\end{bmatrix}dt \tag{11}$$

where $\boldsymbol{x}_i$, $i$=1,2,3 are components of the state vector, $k$ = 1/100 is the reservoir constant, and $F$ is the model forcing.

### 4.1.2 Non-linear model

The only way in which the non-linear model used in the following differs from the linear model is that the constant $k$ for the first two reservoirs is dependent on the state value, thereby creating two different model domains:

$$d\begin{bmatrix}\boldsymbol{x_1}\\ \boldsymbol{x_2}\\ \boldsymbol{x_3}\end{bmatrix}=\begin{bmatrix}F-k_1\boldsymbol{x_1}\\ k_1\boldsymbol{x_1}-k_2\boldsymbol{x_2}\\ k_2\boldsymbol{x_2}-k_3\boldsymbol{x_3}\end{bmatrix}dt \tag{12}$$

where $k_3$ = 1/100 and $k_j=\frac{1}{100} if\ \boldsymbol{x_j}<125\ \ otherwise\ k_j=\frac{1}{200}$, j=1,2;

## 4.2 Observations

The measurement equation for both models is:

$$d=\boldsymbol{Hx}+\epsilon \tag{13}$$

where $d$ is a measurement, the measurement operator $\boldsymbol{H}=[\boldsymbol{0}\ \boldsymbol{0}\ \boldsymbol{1}]$ and ϵ is Gaussian noise representing the observation error. This is used for creating artificial observations as well as for the Kalman filter analysis.

The partial update is tested on both models using gauges with the following four observable ranges:

- >75

- >150
- 75-125
- 95-105

The ranges are illustrated to the right on Figure 5, which also shows the state value of all three reservoirs for a single deterministic simulation with the linear model without any updating. The red curve indicates the state value in the third reservoir (where the gauge is situated). When using a gauge with a lower observation limit of 75, the state is observed the vast majority of the time while the state is only observed for a small fraction of the time when the limit is at 150. The two closed observation intervals are also shown on the figure (the red and green bars). These cover only a limited portion of the range of the variable. The intervals are implemented as gauges that have a lower observation limit and furthermore become saturated when the observed quantity is above the upper observation limit. This means that when the quantity is outside the observed range, the OR-observation is a measurement of the quantity being below or above the observed range, as would be the case for e.g. many concentration measurements. When the OR-observation is below the observed interval, partial updating is applied as when having a lower limit only, while an upper limit implementation is used when the OR-observation is above the interval. Note that the 95-105 interval is so narrow that the observations are almost entirely reduced to being Boolean as 'lower-than' or 'higher-than'.

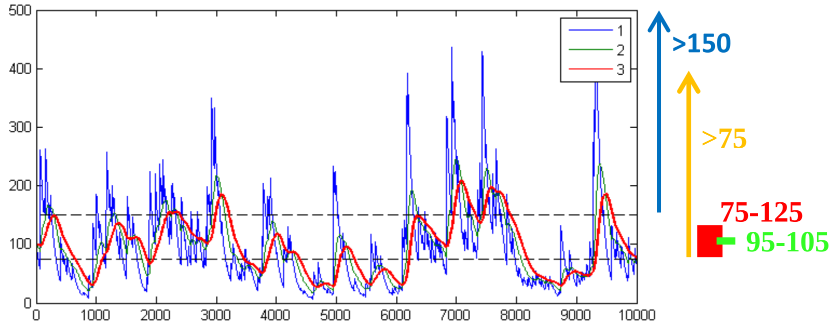

Figure 5: The result of a single deterministic model simulation. The different coloured lines show the values for each of the three state variables. The dotted horizontal lines show two of the lower thresholds used in the following at 75 and 150, respectively. To the right the four gauge observable ranges are displayed.

## 4.3 Forecasting and performance quantification

The tests are performed using an artificial truth created with the same model as used in the tests. Observations are created for the third state variable as perturbations from the truth using Gaussian observation noise with zero mean and a variance of 1. In the same way measured model forcing is created as perturbations of the model forcing used for the truth model by multiplying the true forcing with a factor uniformly distributed between 0 and 2. The same noise description is used for creating forcing perturbations from the measured forcing, to be used for the EnKF. The assumed model noise used with the EnKF is state proportional white Gaussian noise with a standard deviation $\sigma_{model}$ of 0.05 times the state value. The noise is truncated at $\pm 3\sigma_{model}$.

For every time step the model is updated using DEnKF from the observations on the third state variable if the observations are within the defined observable range of the gauge. If the observations are outside the observable range, the partial update scheme is used. For each time step the mean of the updated ensemble is used as basis for producing deterministic forecasts with time horizons up to 300 time steps. The total simulation is $10^4$ time steps long but the first 1000 time steps are used as initialization period and not included in the evaluation. The measured model forcing is used during the forecasts, which means that there are no additional errors related to the forecast of the forcing. Each setup is run 100 times with different realisations of observations and model forcing. The update performance is quantified as the mean of the forecast performance for these 100 simulations.

The main measure of performance is an analogue to the coefficient of determination known as the Nash-Sutcliffe efficiency index $R^2$, (Nash and Sutcliffe, 1970):

$$R^2 = 1 - \frac{SS_{err}}{SS_{total}} = 1 - \frac{\sum_{t=1}^{T}(\boldsymbol{x}_3(t) - \boldsymbol{x\_true_3}(t))^2}{\sum_{t=1}^{T}(\boldsymbol{x\_true_3}(t) - \overline{\boldsymbol{x\_true_3}})^2} \qquad (14)$$

where $T$ is the total number of time steps in the evaluation period, $\boldsymbol{x_3}$ is the forecast initiated from the mean of the updated ensemble, and ***x_true*** is the computed truth. Here, $R^2$ has been preferred over RMSE since the latter is very sensitive to the absolute values of the largest variations, meaning that the overall mean performance of the 100 runs could end up being representative of only a few of the runs. This will not be the case with $R^2$ since this is scaled with the total sum of squares of the individual runs. When $R^2$ is 1, the model predictions are perfect while the mean of the observations is a better predictor than the model when $R^2$ is negative.

The second measure of performance is the median absolute error of each forecast time series which is computed in order to have a measure of performance that is insensitive to the most extreme values but foremost quantifies the typical deviation from the truth.

$$median\ absolute\ error\ = median(|\boldsymbol{x}_3 - \boldsymbol{x\_true_3}|) \qquad (15)$$

Ensemble based data assimilation methods are often used for very computationally expensive models and therefore the required ensemble size is a critical parameter that can determine whether it is at all feasible to use data assimilation with a given model. Therefore the tests are run with both an ensemble size of 100, that would generally be regarded as sufficient, as well as an ensemble size of just 10, in order to investigate how the partial update performs under these different circumstances and to evaluate to which extent the partial update can open up for the use of a much smaller ensemble size.

## 5 Results and Discussion

In the following the prefix “P” is used to indicate that partial updating has been used, so that PDEnKF means that the ensemble was updated with DEnKF when actual observations were available and partial updating has been used in case of OR-observations. Figure 6 shows how much the model forecast skills are improved by PDEnKF when using gauges with various lower limits. The figure shows that none of the two models are very useful as forecast models without updating (green line) since the $R^2$ values in this case are as low as 0.15 and 0 for the linear and non-linear model, respectively. For both models an ensemble size of 10 is clearly too small if the models are to be used to create long-range predictions, since updating makes the models perform worse. In real life applications this would be counteracted by using localization (techniques for limiting the impact of the updates as a function of the distance to the observation (Hamill et al., 2001)), but this has not been used in the following since the purpose of this study is to investigate the isolated effect of the partial updating.

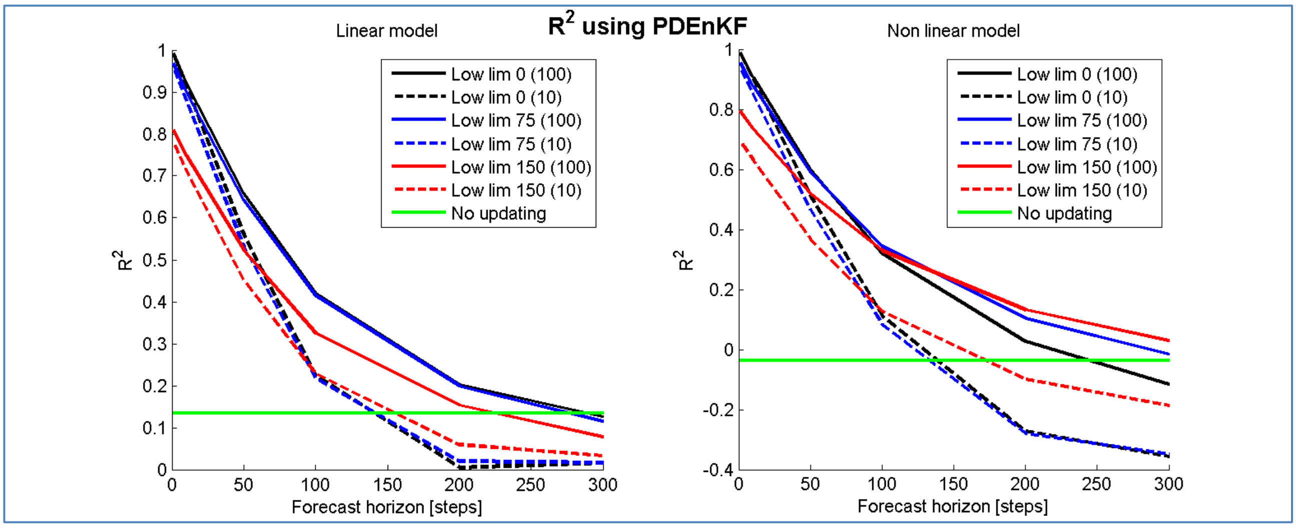

Figure 6: $R^2$ when using partial updating with DEnKF and gauges with various lower limits. The black line is without a lower limit and thereby the partial update has not been used. The green line is the results without updating. Dotted and solid lines indicate ensemble sizes of 10 and 100, respectively.

Interestingly, there seems to be almost no difference between having a lower limit at 75 or at 0 (using DEnKF implementation where all values are assimilated). This suggests that in cases where it is difficult to describe the observation uncertainty for the very low values or where the model is known to show a non-physical behaviour for the low ranges, the partial update can be used to exclude this range from the DA operation without loss of forecasts accuracy. It can be seen from the following figures that a lower limit of 75 leads to some drop in the forecast accuracy if not using partial updating. Note that for the long lead times for the non-linear model it even turns out to be beneficial to use the highest lower limit at 150. This shows that it in some cases is an advantage to disregard observations and just restrict the ensemble spread to some relevant interval.

Figure 7 and Figure 8 compare the performance of DEnKF with and without partial updating (DEnKF vs. PDEnKF). The results consistently show that it is beneficial to use partial updating. As would be expected, the benefit compared to standard DEnKF is very small when the observed quantity is within the observable range most of the time (lower limit at 75) while the improvement is significantly larger when this is not the case. In many of the setups the positive impact on the short to medium range forecasts of introducing partial update when using an ensemble size of 10 is larger than that of increasing the ensemble size to 100. For the non-linear model this is the case even for the longest forecasts when having a lower observation limit at 150. When using the non-linear model with an ensemble size of 10 and the narrow observation interval from 95 to 105, the partial update makes the difference between having a model that does not even produce useful 1 step predictions and having a decent model with some predictive ability up to some hundred steps into the future.

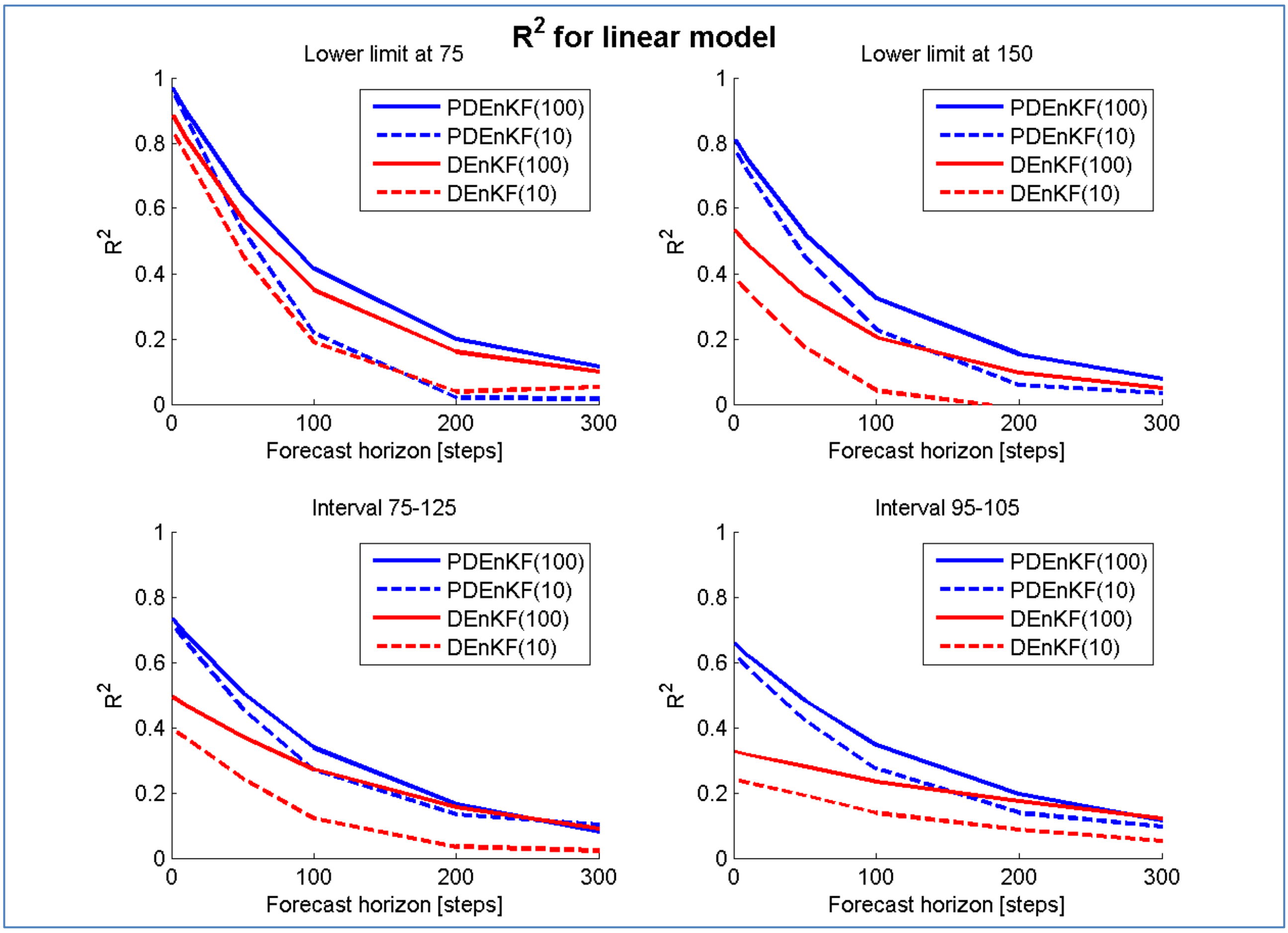

**Figure 7: $R^2$ when using the linear model. The dotted and solid lines show the results when using 10 and 100 ensemble members, respectively. The blue and red lines are with and without partial updating, respectively.**

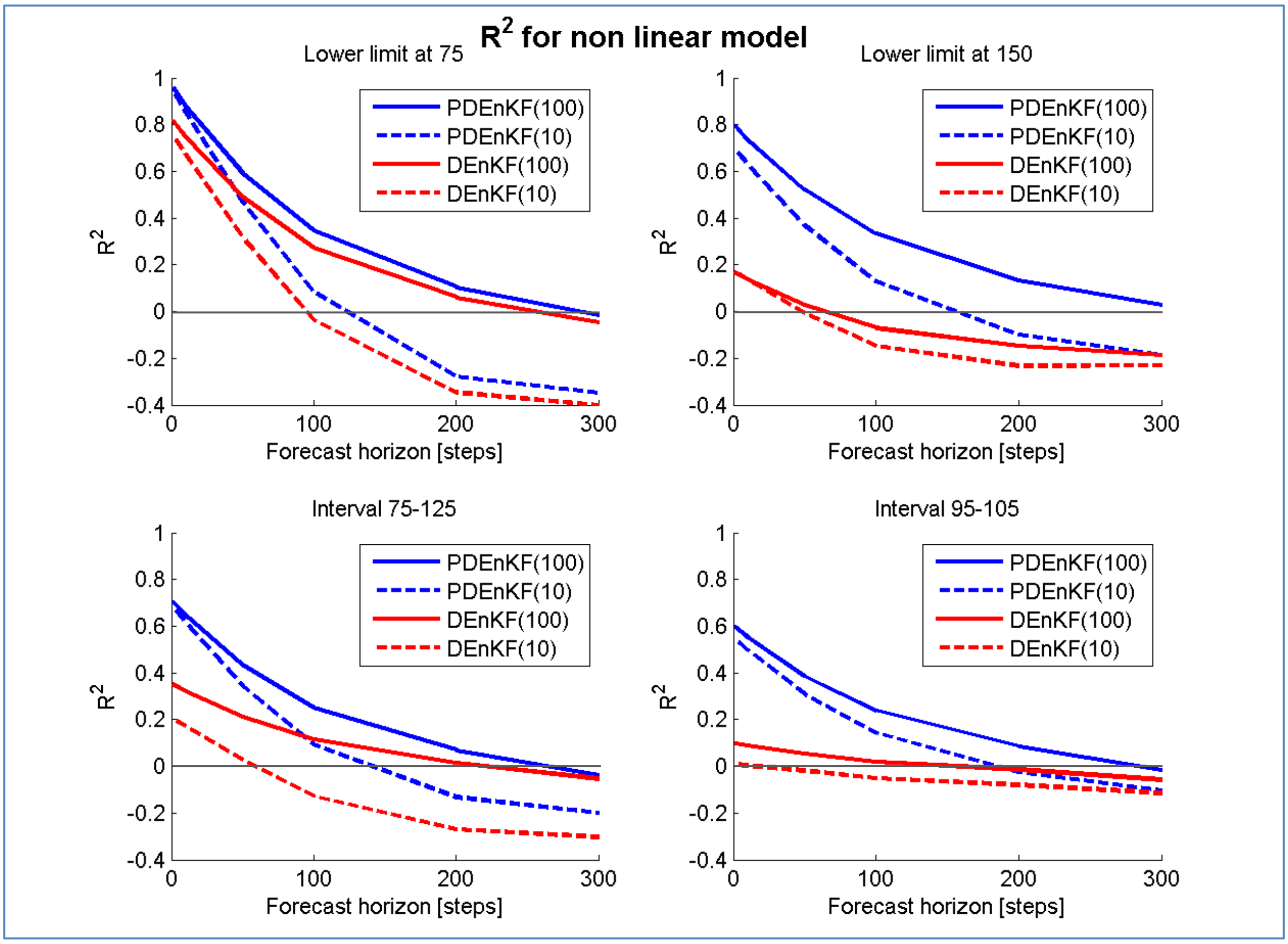

**Figure 8: $R^2$ when using the non-linear model. The dotted and solid lines show the results when using 10 and 100 ensemble members, respectively. The blue and red lines are with and without partial updating, respectively.**

The limits of the $R^2$ values show that the estimation and forecasts of the highest values are improved by the partial update, but since the extremes are always above the lower observation limits the forecasts will often have been initiated while actual measurements were available and are therefore produced under quite different circumstances than the typical forecast. This is in particular true when the lower gauge limit is high. Therefore a plot is shown of the Median Absolute Error for both models using a lower limit of 150, see Figure 9. This reveals that the median error is improved significantly by the partial update even though this performance measure foremost relates to values that are far below the lower observation limit. The ensemble size, on the other hand, has almost no impact on the results.

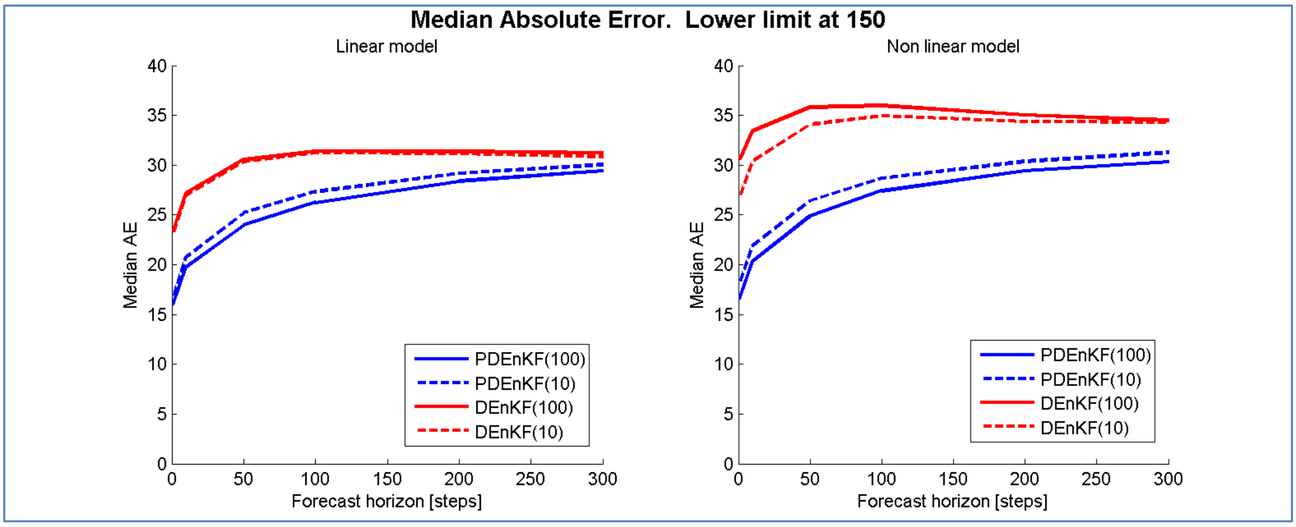

Figure 9: The median absolute error for both models when using a gauge with a lower limit at 150.

All the results shown are based on the DEnKF with and without partial updating. The method can also be used in combination with the standard EnKF formulation with perturbed observations, as long as the partial update is still performed using the deterministic formulation. This was also tested with much the same results as the ones shown here, but with generally lower $R^2$ values for the small ensemble sizes. For clarity, these results have been omitted from the article.

The method is based on using a very non-Gaussian data likelihood for OR-observations. Even though the EnKF assumes strictly Gaussian data likelihoods and is known to suffer from poor performance when this is not the case (Sætrom and Omre, 2011), it was shown that the non-Gaussian likelihood function chosen here could be treated in a consistent way in the EnKF settings. The method can only be justified with the specific OR-observation likelihood function used, which assumes constant likelihood outside the observable range. One could argue that the distribution of an unobserved quantity is not uniform. This is, however, not information related directly to the lack of observations, but rather to the model dynamics, and should therefore not be accounted for by the likelihood function.

If the lack of observations is due to malfunctioning measuring equipment and not due to the observed quantity being outside the observable range, the partial updating scheme will deteriorate the model estimates. The method is very vulnerable to this kind of error and therefore some sort of automated quality control is likely to be required for most real life applications of the method.

When using any data assimilation scheme on large distributed models, there will usually be a need for localization when the ensemble size is superseded by the number of state variables in the model (Oke et al., 2007; Petrie and Dance, 2010). The fact that the OR-observations are treated in the same way as actual observations permits the use of a standard Schur product based localization.

## 6 Conclusions

We propose a method, referred to as *partial updating*, that adds to the versatility of the EnKF by making it possible to utilize the information present in the signal from a gauge when the observed quantity is not within the observable range. The method can be used in the case where a gauge has an upper and/or lower observation limit and therefore is not capable of observing the quantity of interest all the time. In the case where the observations show that the quantity is outside the observable interval but some of the members of the ensemble are inside the interval, these are corrected towards the limit of the interval. The method provides a way of restricting the ensemble spread in periods without actual observations and thereby improving the update of the filter once actual measurements become available, but also improving the immediate state estimate. The results show that it is always beneficial to use partial updating if the gauge has a limited range. The greatest improvement is achieved for a non-linear model. In the most extreme case where the observation interval is very narrow and the observations therefore most of the time is just an indication of whether the quantity is above or below the observation interval, the use of partial updating is absolutely critical for producing skilful forecasts. In many of the tests the positive impact of including partial updating in the EnKF setup greatly outweighs that of increasing the ensemble size from 10 to 100. This shows that the partial updating makes it possible to utilize information that would otherwise be inaccessible for the EnKF.

### Acknowledgements

The research was financially supported by the Danish Council for Strategic Research, Programme Commission on Sustainable Energy and Environment, as part of the Storm and Wastewater Informatics (SWI) project (http://www.swi.env.dtu.dk).

## References

Aubert, D., Loumagne, C., Oudin, L., 2003. Sequential assimilation of soil moisture and streamflow data in a conceptual rainfall–runoff model. Journal of Hydrology 280, 145–161.

Birkel, C., Tetzlaff, D., Dunn, S.M., Soulsby, C., 2010. Towards a simple dynamic process conceptualization in rainfall--runoff models using multi-criteria calibration and tracers in temperate, upland catchments. Hydrological Processes 24, 260–275.

Burgers, G., Jan van Leeuwen, P., Evensen, G., 1998. Analysis Scheme in the Ensemble Kalman Filter. Monthly Weather Review 126, 1719–1724.

Evensen, G., 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. Journal of Geophysical Research 99, 10143–10162.

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

Hamill, T., Whitaker, J., Snyder, C., 2001. Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. Monthly Weather Review 129, 2776–2790.

Houtekamer, P., Mitchell, H., 2001. A sequential ensemble Kalman filter for atmospheric data assimilation. Monthly Weather Review 129, 123–137.

Houtekamer, P.L.P., Mitchell, H.L.H., 1998. Data assimilation using an ensemble Kalman filter technique. Monthly Weather Review 126, 796–811.

Kaleschke, L., Maaß, N., Haas, C., Hendricks, S., Heygster, G., Tonboe, R.T., 2010. A sea-ice thickness retrieval model for 1.4 GHz radiometry and application to airborne measurements over low salinity sea-ice. The Cryosphere 4, 583–592.

Kalman, R.E., 1960. A new approach to linear filtering and prediction problems. Journal of basic Engineering 82, 35–45.

Keppenne, C.L., Rienecker, M.M., 2002. Initial testing of a massively parallel ensemble Kalman filter with the Poseidon isopycnal ocean general circulation model. Monthly weather review 130, 2951–2965.

Lee, J.H., Timmermans, J., Su, Z., Mancini, M., 2012. Calibration of aerodynamic roughness over the Tibetan Plateau with Ensemble Kalman Filter analysed heat flux. Hydrology and Earth System Sciences 16, 4291–4302.

Löwe, R., Mikkelsen, P.S., Madsen, H., 2013. Stochastic rainfall-runoff forecasting: parameter estimation, multi-step prediction, and evaluation of overflow risk. Stochastic Environmental Research and Risk Assessment.

Nash, Je., Sutcliffe, J. V, 1970. River flow forecasting through conceptual models part I—A discussion of principles. Journal of hydrology 10, 282–290.

Nævdal, G., Johnsen, L.M., Hydro, N., Aanonsen, S.I., 2003. SPE 84372 Reservoir Monitoring and Continuous Model Updating Using Ensemble Kalman Filter.

Oke, P.R., Sakov, P., Corney, S.P., 2007. Impacts of localisation in the EnKF and EnOI: experiments with a small model. Ocean Dynamics 57, 32–45.

Olume, V., 2006. Real-Time Data Assimilation for Operational Ensemble Streamflow Forecasting. Journal of Hydrometeorology 7, 548–565.

Petrie, R.E., Dance, S.L., 2010. Ensemble-based data assimilation and the localisation problem. Weather 65, 65–69.

Sakov, P., Evensen, G., Bertino, L., 2010. Asynchronous data assimilation with the EnKF. Tellus A 62, 24–29.

Sakov, P., Oke, P.R., 2008. A deterministic formulation of the ensemble Kalman filter: an alternative to ensemble square root filters. Tellus A 60, 361–371.

Sætrom, J., Omre, H., 2011. Ensemble Kalman filtering for non-linear likelihood models using kernel-shrinkage regression techniques. Computational Geosciences 1–16.

Thorndahl, S., Beven, K.J., Jensen, J.B., Schaarup-Jensen, K., 2008. Event based uncertainty assessment in urban drainage modelling, applying the GLUE methodology. Journal of Hydrology 357, 421–437.

Thorndahl, S., Poulsen, T.S., Bøvith, T., Borup, M., Ahm, M., Nielsen, J.E., Grum, M., Rasmussen, M.R., Gill, R., Mikkelsen, P.S., 2013. Comparison of short-term rainfall forecasts for model-based flow prediction in urban drainage systems. Water science and technology : a journal of the International Association on Water Pollution Research 68, 472–478.

Tong, J., Hu, B., Yang, J., 2012. Assimilating transient groundwater flow data via a localized ensemble Kalman filter to calibrate a heterogeneous conductivity field. Stochastic Environmental Research and Risk … 467–478.

Van Leeuwen, P.J., Evensen, G., 1996. Data Assimilation and Inverse Methods in Terms of a Probabilistic Formulation. Monthly Weather Review 124, 2898–2913.

Whitaker, J., Hamill, T., 2002. Ensemble data assimilation without perturbed observations. Monthly Weather Review 130, 1913–1924.