Stochastic data assimilation of observations with a detection limit

Abhishek Shah

Thesis for the degree of Philosophiae Doctor (PhD)
University of Bergen, Norway
2019

UNIVERSITY OF BERGEN

Stochastic data assimilation of observations with a detection limit

Abhishek Shah

Thesis for the degree of Philosophiae Doctor (PhD)
at the University of Bergen

Date of defense: 16.09.2019

Year: 2019

Title: Stochastic data assimilation of observations with a detection limit

Name: Abhishek Shah

Print: Skipnes Kommunikasjon / University of Bergen

Preface

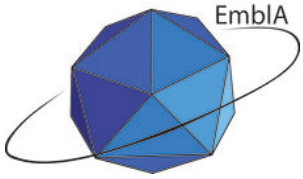
This dissertation is submitted in partial fulfillment of the degree Philosophiae Doctor in mathematics and statistics at the Department of Mathematics, University of Bergen. This work was carried out at the Nansen Environmental and Remote Sensing Center (NERSC), Bergen within the Nordic Center of Excellence EmblA under the supervision of Laurent Bertino and Mohamad El Gharamti.

This thesis includes three research articles describing methodological developments of an ensemble-based data assimilation method that is able to take benefit from observations with a detection limit. The proposed data assimilation method can be specifically used for assimilating sea ice thickness observations that are insensitive to sea ice thicker than a given threshold. The method is developed and tested with various benchmarks: linear and nonlinear models, low-dimensional toy models and a high-dimensional complex dynamical model, using twin and fraternal twin experimental setups. The thesis is structured in two parts. Part I contains a general background, the motivation, contribution, the summary of papers and further perspectives. Part II includes the research articles which are listed below:

Paper I: Assimilation of semi-qualitative observations with a stochastic ensemble Kalman filter. **Shah, A., El Gharamti, M., and Bertino, L.**, published in Quarterly Journal of the Royal Meteorological Society.

Paper II: Assimilation of semi-qualitative sea ice thickness data with the EnKF-SQ. **Shah, A., Bertino, L., Counillon, F., El Gharamti, M., and Xie, J.**, submitted to Tellus A: Dynamic Meteorology and Oceanography.

Paper III: An adaptive correction algorithm for the out-of-range observation error variance of the EnKF-SQ. **Shah, A., El Gharamti, M., Bertino, L., and Counillon, F.**, to be submitted.



Acknowledgements

At last, I am done and a number of people have made this possible. First, I would like to thank my main supervisor Laurent Bertino and co-supervisor Mohamad El Gharamti (aka Moha), for excellent supervision, advice, coaching, guidance, counselling, tutoring and other synonyms. It was a great learning experience for me to work closely with both of them. I would also like to thank my other co-supervisors, Hans Skaug and Geir Evensen for their support. Thanks to François Counillon for his constructive scientific input to my work and for the constant motivation that I can do it!

Many thanks to Patrick, Colin, Yiguo, Jiping, Alberto and Madlen for their occasional illuminating scientific discussions, which have been very helpful in understanding various data assimilation concepts. Thanks to Morten Borup for hosting me at the DTU, Denmark for short a research stay, which has proven to be very important in developing my work. I would like to thank my cheering PhD office mates who always made sure the mood remains light in a serious research atmosphere, specially Michael Hart-Davis. Special thanks to NERSC administration and NERSC's IT department for their friendly conversations and technical support. Thanks to the Nansen Scientific Society and Copernicus Marine Services for funding support.

Thanks to Bergen Gujarati community and dear Åsane gang for making me feel home by regularly organizing small get-together and celebrating festivals.

Special thanks to two important ladies my mom Manju and wife Hetal for their invaluable support, without that I would not have made it. They both have been constant source of inspiration to me. Thanks for being there. Thanks to papa, Akash and the entire Kanku Chopda family for supporting me from far away in India. Finally, thanks to all of them who one or the other way supported me through in this journey of the PhD.

Contents

Part I

Background

Chapter 1

Introduction

1.1 Predictions

A prediction can generally be defined as a guess of what might happen in the future based on recent observations. Numerical weather prediction (NWP) is a popular example that employs a set of equations, which describe the dynamics of the atmosphere and may include those of the ocean and sea ice. In essence, NWP is an initial value problem of mathematical physics, where the future weather state is determined by integrating the governing nonlinear dynamical equations, starting from their observed current state. Because of the nonlinear and chaotic nature of these governing equations, a small discrepancy in the initial conditions can lead to a totally different forecast. Accurate predictions have numerous socio-economic benefits including effective management of energy resources, improved natural disaster planning, mitigation of the impacts from extreme weather events, financial revenues and cost savings in aviation, agriculture and transport, among many others (Shapiro *et al.*, 2009; Williamson *et al.*, 2002). Therefore, it is important to have the best possible estimate of the initial condition in order to obtain an accurate prediction.

The advent of computer simulations in the 1950s opened a new era for predictions. Thereafter came continuous improvements, the introduction of various data assimilation techniques, more efficient numerical models, an increasing availability in-situ observations, satellite data and powerful computational resources. Those have progressively contributed to improving weather predictions. In recent times, climate change and the need for accurate predictions to deal with future extreme events have attracted even more interest and investment in climate research, super-computing capabilities, earth observing satellite missions and other observational programs. The Arctic sea ice has been recognized as a necessary focal point.

The scientific and technological developments during the last $4-5$ decades have improved

the forecast skills significantly. Bauer *et al.* (2015) has mentioned that "forecast skill in the range from 3 to 10 days ahead has been increasing by about one day per decade: today's 6-day forecast is as accurate as the 5-day forecast ten years ago".

This dissertation addresses the use of observations within the framework of data assimilation for improving the accuracy of predictions, in particular on how to make better use of observations with a detection limit while preserving the dynamical consistency and the reliability of predictions.

1.2 Observations

Observations are essential in both environmental and climate science and especially vital for the purpose of statistical inference and estimation. Observational data help infer the underlying distribution of the hidden variables of interest, which cannot be observed directly. For example, a satellite radiometer measures the radiative flux emitted from Earth to outer space at different wavelengths, which are then used to derive the atmospheric temperature and humidity fields through radiative transfer equation (Reale *et al.*, 2008, and references therein). In this typical example of remote sensing observations, atmospheric temperature and humidity fields are hidden states, which are not observed directly but are retrieved indirectly via a functional relation to the observed radiative fluxes. Direct measurements of the variable of interest are more commonly practiced with in-situ measurements rather than satellite measurements: measuring temperature from a thermometer or water level with a tide gauge are among the simplest examples.

1.2.1 Types of observations

An observation of a spatially-distributed variable can generally be categorized into three different types depending on the nature of the data recorded, which are as follows:

1. *Fully quantitative observation:* Observational data, which are available in numerical or quantifiable form. For example, temperatures measured by thermometers, atmospheric radiance profiles measured by satellites, heights of the students in a class among others are fully quantitative data. Figure 1.1 shows a global map of a fully quantitative observational data representing the daily mean atmospheric temperature profile at 850 hPa retrieved by NASA's earth observing system (EOS) Terra satellite.

2. *Fully qualitative observation:* Observational data that are not quantifiable but categorical. For instance the daily sea ice type classification provided by the Exploitation

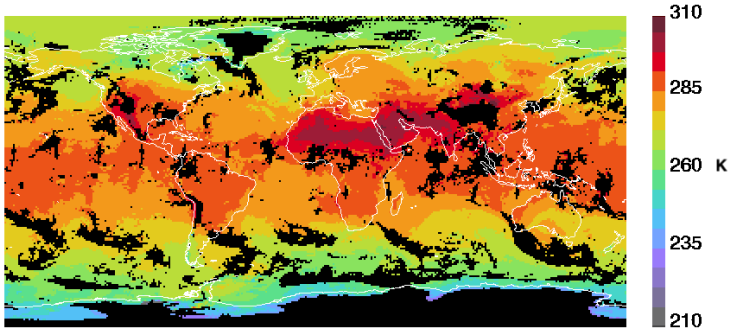


Figure 1.1: Global map of the mean atmospheric temperature profile on 06 June 2019 retrieved by the EOS TERRA. The map is obtained from https://modis-images.gsfc.nasa.gov.

of Meteorological Satellites (EUMETSAT) Ocean and Sea Ice Satellite Application Facilities (OSI-SAF) (Aaboe *et al.*, 2018), is a fully qualitative data type. The sea ice type classification (open water, first-year ice, multi-year ice) is based on the analysis passive microwave and scatterometry data over the entire Arctic Ocean. Figure 1.2 shows daily sea ice type classification as a qualitative observational data from the OSI-SAF. The previous example is originally a quantitative observation (emissivity or roughness of the ocean or ice surface) that cannot be exploited as a geophysical measurement but can be converted to a qualitative observation by a classification algorithm. Further examples are quite common in geosciences such as soil occupation types, vegetation classification and cloud masking. In studies related to the remediation of polluted soils, the presence of bad smell can also be used as qualitative data.

3. *Semi-qualitative observation:* Observational data, which are partly quantitative and qualitative in nature are defined as semi-qualitative. This type of observations arise primarily because of a detection limit in the measuring instrument. Even though quantitative data cannot be recorded outside the observing range of the instrument, a qualitative indication that the observed quantity is above or below the detection limit is available. Some examples of semi-qualitative observations are the contaminant concentrations with lower detection limit in environmental and health fields (Hornung and Reed, 1990), water levels in urban water networks (Borup *et al.*, 2015), river water level measurements with lower detection limit of 1 km obtained from satellite radar altimetry (Birkett, 1998), Soil Moisture Ocean Salinity (SMOS) satellite retrieved sea ice thickness (SIT) with upper detection limit of 50 cm (Kaleschke *et al.*, 2012) and thick sea ice thickness obtained from the CryoSAT2 satellite with a lower detection limit (Laxon *et al.*, 2013). Retrievals of sea ice concentrations from passive microwave remote sensing are also limited in range by the "weather filter" (Ivanova *et al.*, 2015).

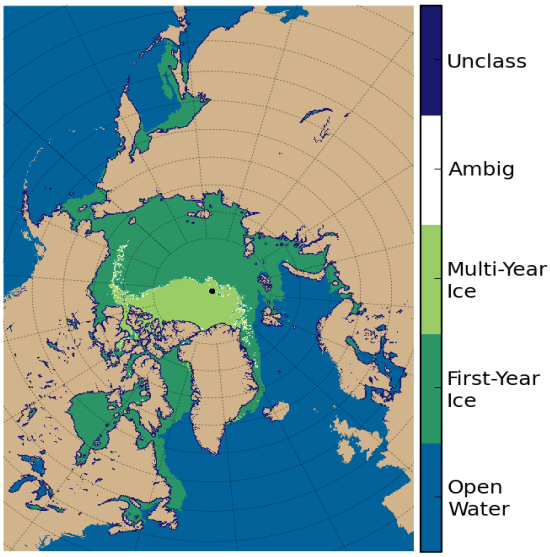


Figure 1.2: The daily sea ice type classification from the OSI-SAF on 05 February 2019. The image is taken from http://osisaf.met.no

Other types of observations that are not covered above include those with discrete densities such as observations of icebergs, whales, asteroids, etc. This PhD thesis only considers variables that are continuously distributed in space and commonly used in climate and environmental forecasting.

In the following dissertation, we focus on semi-qualitative observations. The goal is to learn how to use such information to estimate and infer the underlying distribution of hidden state variables. Article II (Shah *et al.*, 2019) for instance presents a case study, where semi-qualitative sea ice thickness observations are used to estimate the model estimate of the same variable. Therefore, in the next section we will briefly go through the SMOS retrieved sea ice thickness.

1.2.2 The case of SMOS sea ice thickness

The SMOS satellite carries onboard a novel interferometric radiometer that operates in 1.4 Ghz L-band microwave range, which can capture *brightness temperature* images. This is an example of passive microwave remote sensing, that is, it only records the natural microwave reflected and emitted by the earth and its atmosphere rather than sending active signals as in active microwave remote sensing. The L-band, in contrast to more commonly used C-band and X-band, contains longer wavelengths that can penetrate deeper in the medium of interest, such as ocean, glaciers or sea ice. Figure 1.3, shows a simple schematic representation of the passive remote sensing of sea ice thickness for the SMOS mission.

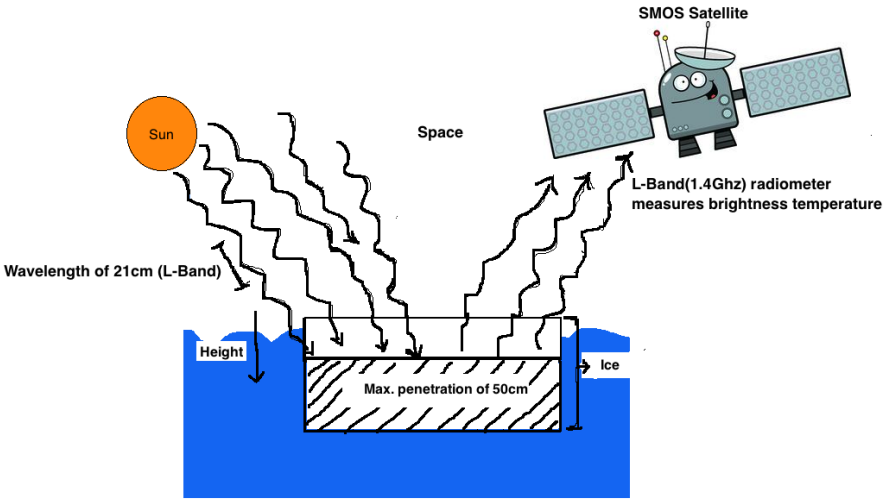


Figure 1.3: Very simple schematic representation of passive remote sensing of the sea ice thickness in the case of the SMOS satellite.

Only the top $\sim$ 50 cm of the sea ice is permeable to L-band microwaves (Kaleschke *et al.*, 2012), and hence the maximum retrievable ice thickness with acceptable uncertainty using SMOS is approximately equal to 50 cm, with some nuances depending on the deformation state of sea ice. Thin sea ice signals, however, are only available in the cold months. In the melting season, the emission properties in the microwave are polluted by the wetness of the surface and occurrence of melt ponds in the Arctic. Therefore, thickness data in the Arctic are calculated only during the freezing season, that is from October to April. During the melting season, the procedure does not yield meaningful results.

The typical behavior of the semi-qualitative SMOS SIT can be seen in Figure 1.4, which shows SMOS derived SIT against a proxy of ice thickness (number of freezing days times the degrees below freezing point). The plot demonstrates that for SMOS SIT estimates thicker than 30 cm, the uncertainties progressively increase and become very high (approx. double than the SIT estimates itself) for thickness $>$ 50 cm. Therefore, SMOS SIT estimates $>$ 50 cm are simply considered qualitative, making the entire SMOS SIT product semi-qualitative with upper detection limit of 50 cm.

A similar satellite mission called SMAP (Soil Moisture Active Passive) has been launched by NASA in 2015, which has also been used for the retrieval of range-limited sea ice thickness (Patilea *et al.*, 2019). Passive microwave measurements in shorter wavelengths follow the same principle but the return signal comes from a much thinner layer of ice and can be considered as "surface measurements" only. An approach to measure sea ice thickness by passive microwave of even longer wavelengths than the L-band has been proposed by Macelloni *et al.* (2018) but not yet selected, in particular due to the expected noise from

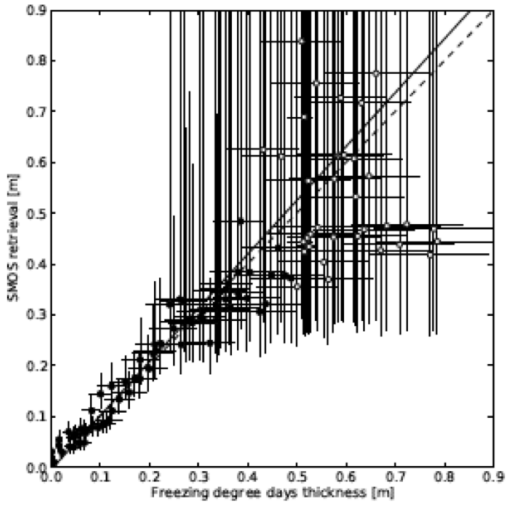


Figure 1.4: SMOS derived sea ice thickness (y-axis) versus a proxy of ice thickness (x-axis). Vertical and horizontal solid lines show the uncertainties in SMOS SIT estimates and proxy of ice thickness, respectively. Credit: Kaleschke, personal communication.

Radio-Frequency Interferences.

The next chapter presents the standard methods and tools used in data assimilation to estimate the underlying state distribution of model variables given these types of observations.

Chapter 2

Methods and Tools

This chapter gives a general introduction of the different approaches to the hidden state-estimation process. It starts with the brief overview of the geostatistical techniques used to infer static hidden state variables in the case of semi-qualitative observations. A more comprehensive overview follows of different data assimilation methods used to infer dynamical state variables.

2.1 Inference of static variables with Geostatistics

Geostatistics is the branch of statistics used to analyze and predict the variables distributed in space and time. Many geostatistical methods were originally developed to estimate the spatial patterns of underlying static hidden state variables and interpolate values for locations where observations were not taken. One such example is the mining industry application, where geostatistical modelling is applied to estimate the mineral grades within the ore deposits (Chiles and Delfiner, 2012; Journel and Huijbregts, 1978). The precise estimation of the mineral grades is of great economical importance to the miners and therefore efficient statistical methods are required. To this end, geostatisticians apply techniques like Kriging to predict the mineral grade at any unobserved location of the ore deposit by accounting for any available neighboring data and their spatial correlations. Kriging predicts the value of a state at a certain location by interpolating values from neighbouring locations and using a weighted average of these known values. To attain useful estimators, strong assumptions are made on the statistical properties of the random fields (random variables distributed in space and/or time), typically that of *second order* stationarity: the expectation of the field and its spatial covariance structure are both invariant in space (and time). Therefore geostatistics are most commonly applied to static variables, which can be sampled repeatedly over long periods of time.

Since the Kriging technique relies on a weighted average of the data values, it yields a too smooth image of the true grade distribution in the deposit and therefore misses the small spatial scales that are not observed properly. Conditional simulations are another family of techniques from geostatistics that are used to generate randomly small scale features between observations while simultaneously honouring the observations by the application of Kriging.

Journel (1986) first addressed the possibility to include semi-qualitative data in Kriging estimators and introduced the vocabulary of "soft" and "hard" data, which we will adopt here. Emery and Robles (2008) proposed techniques for the geostatistical simulation of mineral grades, which can efficiently include soft data. Related approaches are reviewed in Chiles and Delfiner (2012) and followed up by Emery *et al.* (2014): These simulation techniques are all based on Markov Chain Monte Carlo (MCMC) sampling with the realizations conditioned on available observations, both hard and soft data. Hard data, for example, are assays on ore samples and soft data the rock-type information. The sampling method employed by Emery and Robles (2008) is an MCMC with a Gibbs sampler to simulate the mineral grades in ore deposits conditioned on hard and soft data in an approach called *interval constraints* (for detailed description, the reader is referred to (Section 2; Emery and Robles, 2008)). Such a technique can effectively make use of the semi-qualitative data to reconstruct the true spatial trends. The computational cost of these iterative MCMC sampling algorithms is a recognized issue when processing large datasets, although these algorithms have already been designed to have a high frequency of acceptance (Marcotte and Allard, 2018).

Next, we study the estimation of a hidden state variable that evolves in time according to nontrivial physical equations. In other words, consider a dynamical hidden state variable that necessitates the resolution of a computationally intensive numerical model. For such cases, the aforementioned second order stationarity assumption breaks down. MCMC methods requiring thousands of cycles to simulate large datasets also become too costly to be used in conjunction with a forward dynamical model. Therefore, realistic and feasible inference techniques for the dynamical hidden state variables are required to converge within the order of a hundred of random realizations. In the following section, we discuss practical approaches for the inference of dynamical hidden state variables in the context of data assimilation.

2.2 Data assimilation

Data assimilation (DA) is an approach for fusing observations with model forecasts to obtain a best possible estimate of the true state of the process together with the associ-

ated uncertainties (Wikle and Berliner, 2007). It is mainly applied for *state estimation* of dynamical systems but has also been used for estimating uncertain model parameters (e.g., Gharamti *et al.*, 2015). The application of DA includes numerical weather prediction (NWP) for forecast initialization (Ghil and Malanotte-Rizzoli, 1991; Lorenc, 1986), history matching in reservoir engineering for oil production and planning (Aanonsen *et al.*, 2009; Oliver and Chen, 2011), atmosphere and ocean reanalysis (Dee *et al.*, 2011; Kalnay *et al.*, 1996; Xie *et al.*, 2017), coupled reanalysis of the earth system (Counillon *et al.*, 2016; Laloyaux *et al.*, 2018), epidemiology analysis and prediction (Pasetto *et al.*, 2017; Rhodes and Hollingsworth, 2009), see a recent review of data assimilation for the geosciences in Carrassi *et al.* (2018). Depending on the application, the number of state variables to be estimated can vary between few hundreds to $\mathcal{O}(10^9)$. The modern day DA approaches can be classified as follows:

- Variational DA methods: In this approach an optimal state trajectory that best fits observational data over a time window is found by minimizing a cost function. The estimated state variable at the end of the time window is then used to initialize the dynamical model for computing the forecast. 3D-Var (Lorenc, 1986) and 4D-Var (Le Dimet and Talagrand, 1986; Lewis and Derber, 1985) are the two most common methods. 4D-Var requires the development and maintenance of an adjoint model. Traditionally, variational DA schemes use a fixed background error covariance matrix unlike it's sequential competitors.

- Sequential DA methods: Here the observations are assimilated in the model every time they become available. The important feature of this kind of methods is the flow-dependant background error covariance, otherwise known as *errors of the day*. The Kalman filter (Kalman *et al.*, 1960) and ensemble Kalman filter (Evensen, 1994; Houtekamer and Mitchell, 1998) among other ensemble-based filters are examples of sequential DA methods.

- Hybrid DA methods: This kind of methods are constructed by combining the variational and sequential DA flavours. The motivation is to make use of a flow-dependent background error covariance matrix in a variational framework. Examples of hybrid DA methods include EnKF-3Dvar (Gharamti *et al.*, 2014; Hamill and Snyder, 2000), 4DEnVar (Buehner *et al.*, 2010a,b; Liu *et al.*, 2008) and iterative ensemble Kalman smoothers (Bocquet and Sakov, 2014), see Carrassi *et al.* (2018) for a review of hybrid methods. All these methods can be used for recursive problems with an assimilation window that moves forward in time.

This thesis makes use of only the sequential DA approach because it allows a simple stochastic treatment of the problem using Monte Carlo (i.e. *ensemble*) techniques. The use of

semi-qualitative data in variational DA has been discussed in Bocquet *et al.* (2010) but is quite discrete in the literature. Applications in hybrid DA methods should in principle be possible, although they may be algorithmically more complex. In the next section, we will introduce the concept of the dynamical hidden state and thereby present a hidden Markov model. The hidden Markov model constitutes the underlying statistical model for DA.

2.2.1 The dynamical hidden state model

In this section, we look at the DA problem from a Bayesian perspective for a process that evolves over time and constrained with physical equations.

A *stochastic process* is a set of $\{\mathbf{x}_k : k \in \mathbb{T}\}$ where k is an index variable belonging to the index set $\mathbb{T}$, and $\mathbf{x}_k$ is a random variable or a vector of several random variables. Here, the index variable k represents time and therefore the index set $\mathbb{T}$ is some subset of $\mathbb{R}$. For spatiotemporal processes, the index variable will be a combination of time and spatial coordinates. In cases when the random variables $\mathbf{x}_k$ of the stochastic process take values in the state space then the underlying statistical model will be called a *state space model*. If the random variables $\mathbf{x}_k$ are discrete in time, then the stochastic process has a *discrete state space* or if it is continuous then the process is said to have a *continuous state space* (Jazwinski, 1970).

Following Jazwinski (1970), the unknown hidden state $\mathbf{x}_k \in \mathbb{R}^n$ and observation $\mathbf{y}_k \in \mathbb{R}^m$ are generated for sequentially increasing time index k, by a dynamical model, $\mathscr{M} : \mathbb{R}^n \to \mathbb{R}^n$, and an observation model, $h : \mathbb{R}^n \to \mathbb{R}^m$, as follows:

$$\mathbf{x}_{k+1} = \mathscr{M}(\mathbf{x}_k) + \mathbf{q}_k, \qquad k = 0, 1, 2, ..., \tag{2.1}$$

$$\mathbf{y}_k = h(\mathbf{x}_k) + \mathbf{r}_k, \qquad k = 1, 2, 3, ..., \tag{2.2}$$

where the Gaussian white noise processes $\mathbf{q}_k$ and $\mathbf{r}_k$, and the initial condition, $\mathbf{x}_0$, are specified by:

$$\mathbf{x}_0 \sim \mathscr{N}(\boldsymbol{\mu}_0, \mathbf{P}_0), \tag{2.3}$$

$$\mathbf{q}_k \sim \mathscr{N}(0, \mathbf{Q}), \tag{2.4}$$

$$\mathbf{r}_k \sim \mathscr{N}(0, \mathbf{R}), \tag{2.5}$$

where $\mathscr{N}(\boldsymbol{\mu}_0, \mathbf{P}_0)$, $\mathscr{N}(0, \mathbf{Q})$ and $\mathscr{N}(0, \mathbf{R})$ are multivariate Gaussian probability distributions with the first and second term inside the brackets representing the mean and covariance respectively. The Gaussian assumption can be relaxed as indicated by Carrassi *et al.* (2018).

Figure 2.1 is a graphical representation of the stochastic process $\{(\mathbf{x}_k, \mathbf{y}_k) : k = 1, 2, \ldots\}$, which constitutes a hidden Markov model (HMM). The HMM is a sequence of the hidden

states $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \ldots$ linked together by a dynamical model equation, such as Eq. 2.1, which are only observed through an observation model, such as Eq.2.2. The horizontal arrows in Figure 2.1 represent the causality (cause and effect) introduced by the dynamical forward model $\mathscr{M}$ and the vertical arrows represent the relationship between the state and the observation of the state for the corresponding time index k.

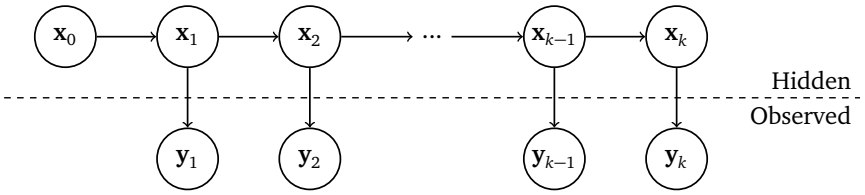


Figure 2.1: Graphical representation of the hidden Markov model.

The underlying stochastic process of the HMM is a Markov process. The Markov process has a Markov property, that is, the conditional probability distribution of future states of the process (conditional on both past and present states) depends only upon the present state, not on the sequence of the events that preceded it. The Markov property is a fundamental property of the HMM representing the conditional independence relation, which is given as follows:

$$p\left(\mathbf{x}_{k+1}|\mathbf{x}_k, \mathbf{x}_{k-1}, \ldots, \mathbf{x}_0\right) = p\left(\mathbf{x}_{k+1}|\mathbf{x}_k\right), \tag{2.6}$$

$$p\left(\mathbf{y}_k|\mathbf{x}_k, \mathbf{x}_{k-1}, \ldots, \mathbf{x}_0\right) = p\left(\mathbf{y}_k|\mathbf{x}_k\right), \tag{2.7}$$

where the probability terms appearing in Eq. 2.6 and 2.7 are the conditional probabilities and $\mathbf{x}_{k-1}, \ldots, \mathbf{x}_0$ is the sequence of the states $\mathbf{x}$. Eq. 2.6 shows the conditional independence of the state at time $k+1$ from all the earlier states except the immediately previous one. Similarly, Eq. 2.7 represents the conditional independence of observation from all the other states except the current one.

2.2.2 Bayesian inference

DA is used to refer to a range of inference procedures whereby observations are assimilated into a statistical model of a dynamical system. In this section, we will formulate the DA problem as statistical sequential inference on a HMM. The main objective of the Bayesian inference is to compute the posterior distribution (also known as *analysis*) $p\left(\mathbf{x}|\mathbf{y}\right)$, of the hidden state $\mathbf{x}$ conditioned on the observation $\mathbf{y}$ using Bayes' rule

$$p\left(\mathbf{x}|\mathbf{y}\right) = \frac{p\left(\mathbf{y}|\mathbf{x}\right)p\left(\mathbf{x}\right)}{p\left(\mathbf{y}\right)} \propto p\left(\mathbf{y}|\mathbf{x}\right)p\left(\mathbf{x}\right), \tag{2.8}$$

where $\mathbf{y}$ is any available data, and $\mathbf{x}$ is the unknown state to be estimated. The probability density function (pdf) $p(\mathbf{x})$ is called a *prior* distribution of $\mathbf{x}$, which quantifies the prior information about it and $p(\mathbf{y}|\mathbf{x})$ is the *observation likelihood* providing information about the data. $p(\mathbf{y})$ is the *marginal* density of the observation $\mathbf{y}$, as considered as a *normalizing constant*.

Let $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K$ and $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_K$ be the sequences of the model states and observations constituting the HMM, within the time interval $[t_0, t_K]$, respectively. In light of the conditional independence property of HMM and the assumption that the observations are independent in time, one can write the following equations for the observation likelihood and the prior pdf:

$$p(\mathbf{y}_{1:K}|\mathbf{x}_{0:K}) = \prod_{k=1}^{K} p(\mathbf{y}_k|\mathbf{x}_k), \tag{2.9}$$

$$p(\mathbf{x}_{0:K}) = p(\mathbf{x}_0) \prod_{k=1}^{K} p(\mathbf{x}_k|\mathbf{x}_{k-1}). \tag{2.10}$$

Applying Bayes' rule on the above two equations, we get a product form for the posterior distribution:

$$p(\mathbf{x}_{0:K}|\mathbf{y}_{1:K}) \propto p(\mathbf{x}_0) \prod_{k=1}^{K} p(\mathbf{y}_k|\mathbf{x}_k)\, p(\mathbf{x}_k|\mathbf{x}_{k-1}). \tag{2.11}$$

Eq. 2.11 suggests that as new data becomes available, one can update the previous optimal estimate of the state process without having to start computation from the beginning. This will create a chain of sequential or recursive updates starting from $p(\mathbf{x}_0)$ and alternating between a prediction (or forecasting) step to obtain the forecast:

$$p(\mathbf{x}_{k+1}|\mathbf{y}_{1:k}) = \int p(\mathbf{x}_{k+1}|\mathbf{x}_k)\, p(\mathbf{x}_k|\mathbf{y}_{1:k})\, d\mathbf{x}_k, \tag{2.12}$$

and an updating (or filtering) step to obtain the analysis:

$$p(\mathbf{x}_{k+1}|\mathbf{y}_{1:k+1}) = \frac{p(\mathbf{x}_{k+1}|\mathbf{y}_{1:k})\, p(\mathbf{y}_{k+1}|\mathbf{x}_{k+1})}{p(\mathbf{y}_{k+1})}. \tag{2.13}$$

The right hand side of Eq. 2.12 is also known as the Chapman-Kolmogorov equation. The alternate application of prediction and update steps as new data becomes available yields the distribution pairs $p(\mathbf{x}_1|\mathbf{y}_1)$; $p(\mathbf{x}_2|\mathbf{y}_1)$, $p(\mathbf{x}_2|\mathbf{y}_{1:2})$; ...; $p(\mathbf{x}_K|\mathbf{y}_{1:K-1})$, $p(\mathbf{x}_K|\mathbf{y}_{1:K})$.

In practice, one may not be able to obtain analytical representations for the forecast and analysis pdfs. However, in the case of Gaussian prior distribution, Gaussian observation error distribution, linear dynamical model and linear observation operator, one can obtain the posterior or analysis explicitly, which will also be Gaussian. The Gaussian and linear case yields the famous Kalman filter.

2.2.3 The Kalman filter

In this section we will briefly introduce the Kalman filter (KF) (Kalman *et al.*, 1960) equations for state estimation without derivation. For the HMM that are Gauss-linear, that is, a linear forward dynamical model and observation operator, Gaussian observations and model error distribution along with Gaussian prior pdf, the Kalman filter (KF) is the optimal solution for sequential updating. The posterior distribution is Gaussian and therefore the pdfs involved are fully characterized by their first two moments: (i) mean and (ii) covariance. Rather than computing a new estimate of the posterior pdfs at every time step, it is sufficient to update the prior mean and covariance with the observations in order to obtain the full posterior pdf.

Let $\mathbf{M}_{k-1:k} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a linear dynamical model integrating the state from time t_{k-1} to t_k and $\mathbf{H}_k : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is a observation operator at time index k. Suppose $p\left(\mathbf{x}_{k-1}|\mathbf{y}_{1:k-1}\right) = \mathcal{N}\left(\mathbf{x}_{k-1}|\mathbf{x}^a_{k-1}, \mathbf{P}^a_{k-1}\right)$, where $\mathbf{x}^a_{k-1}$ and $\mathbf{P}^a_{k-1}$ are the analysis mean and covariance matrix for the corresponding time index, $k-1$. Then the prediction Eq. 2.12 gives $p\left(\mathbf{x}_k|\mathbf{y}_{1:k-1}\right) = \mathcal{N}\left(\mathbf{x}_k|\mathbf{x}^f_k, \mathbf{P}^f_k\right)$ and the corresponding KF forecast equation for the mean and covariance are as follows:

$$\mathbf{x}^f_k = \mathbf{M}_{k-1:k}\mathbf{x}^a_{k-1}, \tag{2.14}$$

$$\mathbf{P}^f_k = \mathbf{M}_{k-1:k}\mathbf{P}^a_{k-1}\mathbf{M}^T_{k-1:k} + \mathbf{Q}_k, \tag{2.15}$$

where the superscripts a, f, and T stand for analysis, forecast and matrix transpose, respectively. Once the forecast step is done, the analysis Eq. 2.13 yields $p\left(\mathbf{x}_k|\mathbf{y}_{1:k}\right) = \mathcal{N}\left(\mathbf{x}_k|\mathbf{x}^a, \mathbf{P}^a\right)$, where the analysis mean $\mathbf{x}^a$ and covariance matrix $\mathbf{P}^a$ are now associated with time index k and the corresponding KF analysis or update equations are given by

$$\mathbf{x}^a_k = \mathbf{x}^f_k + \mathbf{K}_k\left(\mathbf{y}_k - \mathbf{H}_k\mathbf{x}^f_k\right), \tag{2.16}$$

$$\mathbf{P}^a_k = \left(\mathbf{I}_k - \mathbf{K_k}\mathbf{H}_k\right)\mathbf{P}^f_k, \tag{2.17}$$

where $\mathbf{I}_k \in \mathbb{R}^{n\times n}$ is the identity matrix, the term $\mathbf{y}_k - \mathbf{H}_k\mathbf{x}^f_k$ is called the *innovations* (as it is the sole entry point of new observations) and the *Kalman gain matrix*, $\mathbf{K} \in \mathbb{R}^{n\times m}$, is

$$\mathbf{K}_k = \mathbf{P}^f_k\mathbf{H}^T_k\left(\mathbf{H}_k\mathbf{P}^f_k\mathbf{H}^T_k + \mathbf{R}_k\right)^{-1}. \tag{2.18}$$

Therefore, the KF consists of repeating the matrix computations of Eq. 2.14 to 2.17 for sequentially increasing k. The resulting analysis state estimate $\mathbf{x}^a_k$, has minimum error variance and is unbiased. The KF only works optimally for linear models, which are rare in practice but its most common nonlinear variant, the extended Kalman filter, was proven to be unstable for chaotic nonlinear dynamics (Evensen, 1992).

2.2.4 The Ensemble Kalman filter

The ensemble Kalman filter (EnKF) (Burgers *et al.*, 1998; Evensen, 1994, 2003) is an ensemble-based variant of the KF. The central idea behind the EnKF is to use a Monte Carlo method to generate an ensemble of model states $[\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N]$, that are independent identically-distributed realizations of the process, to sample the state space. The positive integer N denotes the ensemble size. In this section we will present the forecast and update step equations of the EnKF without going into detailed derivations.

Conceptually, the EnKF only differs from the KF by the way it treats the propagation step, that is, each ensemble member is integrated with the fully nonlinear model $\mathscr{M}$ from one observation time to the next. The evolution for each $\mathbf{x}_i^a$, where $i \in [1, N]$, is thus computed using $\mathscr{M}$ as follows (hereafter, for clarity, the time index is omitted from the notations):

$$\mathbf{x}_i^f = \mathscr{M}\left(\mathbf{x}_i^a\right) + \mathbf{q}_i, \tag{2.19}$$

where subscript i represents the ith ensemble member. The forecast ensemble can be used to estimate the ensemble forecast error covariance matrix as

$$\hat{\mathbf{P}}^f = \frac{1}{N-1}\sum_{i=1}^{N}\left(\mathbf{x}_i^f - \bar{\mathbf{x}}^f\right)\left(\mathbf{x}_i^f - \bar{\mathbf{x}}^f\right)^T, \tag{2.20}$$

where $\bar{\mathbf{x}}^f$ is the mean of the forecast ensemble

$$\bar{\mathbf{x}}^f = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i^f. \tag{2.21}$$

Once the forecast ensemble is computed, available observations $\mathbf{y} \in \mathbb{R}^m$ are used to compute the updated analysis ensemble. Each forecast member is updated individually using the KF update equation given by Eq. 2.16. The update step of the EnKF is given as follows:

$$\mathbf{x}_i^a = \mathbf{x}_i^f + \hat{\mathbf{K}}\left(\mathbf{y}_i - \mathbf{H}\mathbf{x}_i^f\right), \tag{2.22}$$

$$\hat{\mathbf{K}} = \hat{\mathbf{P}}^f\mathbf{H}^T\left(\mathbf{H}\hat{\mathbf{P}}^f\mathbf{H}^T + \mathbf{R}\right)^{-1}, \tag{2.23}$$

where $\mathbf{y}_i$ is is the ith perturbed observation vector sampled from a Gaussian distribution $\mathcal{N}(\mathbf{y}, \mathbf{R})$; $\hat{\mathbf{K}}$ is the Kalman gain matrix computed using the ensemble statistics. Eq. 2.22 is analogous to the aforementioned process of geostatistical conditional simulations, but this time it makes use of a dynamical model to generate the forecast (Chiles and Delfiner, 2012). For simplicity we assumed a linear observation operator $\mathbf{H}$ in the EnKF update Eq. 2.22 and 2.23 but the EnKF can be applied for nonlinear observation operators as well (Evensen, 2003). Similar to Eq. 2.20 the analysis-error covariance matrix can be computed from the

ensemble of analysis states. Under Gauss-linear assumptions, $\hat{\mathbf{K}}$ converges to the optimal Kalman gain given in Eq. 2.18 as $N \to \infty$, and the distribution of the analysis ensemble $\left[\mathbf{x}_1^a, \mathbf{x}_2^a, ..., \mathbf{x}_N^a\right]$ converges to the true posterior distribution (Le Gland *et al.*, 2009). For non-Gaussian and nonlinear scenario there are so far no general convergence results but still the EnKF has shown to perform well in many such DA problems (Asch *et al.*, 2016).

2.2.5 The Deterministic Ensemble Kalman filter

The stochastic EnKF employs an observation perturbation strategy for updating individual forecast ensemble members. The perturbation of observations is necessary to match the analysis covariance of the KF, otherwise the variance will be systematically underestimated. This is, however, an additional source of sampling error and can become problematic for small ensemble sizes. To counteract the issue of sampling error several *deterministic* square root ensemble-based DA schemes have been proposed, which compute the analysis without perturbing the observations and also preserve the KF posterior covariance during the update. Examples of the square root ensemble schemes are the ensemble transform Kalman filter (ETKF; Bishop *et al.*, 2001; Whitaker and Hamill, 2002), the ensemble adjustment Kalman filter (EAKF; Anderson, 2001), square root analysis scheme for EnKF by Evensen (2004) and the deterministic ensemble Kalman filter (DEnKF; Sakov and Oke, 2008).

The DEnKF was proposed as a simple modification of the ETKF, which results in an asymptotic matching of the analysed error covariance given by the KF in cases where the analysis increments are relatively small. The simple modification uses a first order Taylor expansion of the ETKF update scheme and tends to over-estimate the ensemble variance: it implicitly inflates the analysis ensemble, which is a desirable feature for small ensembles (Raanes *et al.*, 2019). The DEnKF requires only marginal changes to available stochastic EnKF code.

The analysis ensemble in the DEnKF is computed as in Sakov and Oke (2008):

- First, the analysis ensemble mean $\bar{\mathbf{x}}^a$ is computed using the classical KF update Eq. 2.16.

- The analysis anomalies $\mathbf{A}^a$ are then calculated

$$\mathbf{A}^a = \mathbf{A}^f - \frac{1}{2}\mathbf{K}\mathbf{H}\mathbf{A}^f, \tag{2.24}$$

  where $\mathbf{A}^f$ is the forecast anomalies matrix, whose columns are the deviations from the ensemble mean; that is, for $i = 1, 2, ..., N$,

$$\left[\mathbf{A}^f\right]_i = \mathbf{x}_i^f - \bar{\mathbf{x}}^f.$$

- The analysis ensemble members are then deduced from the mean and the anomalies:

$$\mathbf{x}_i^a = \bar{\mathbf{x}}^a + [\mathbf{A}^a]_i \,, \tag{2.25}$$

  where $[\mathbf{A}^a]_i$ is the ith column of the analysis anomalies matrix.

In essence, the analysis scheme of the DEnKF is equivalent to applying the EnKF update equation to each anomaly using half the Kalman gain and without perturbing the observations.

2.2.6 The Partial Deterministic Ensemble Kalman filter

Until now, the presented ensemble-based DA methods only assimilate hard data, that is, quantitative observations. To the best of our knowledge only one study (Borup *et al.*, 2015), has dealt with the issue of observations with detection limit in an ensemble-based DA framework.

Borup *et al.* (2015) has proposed the partial deterministic ensemble Kalman filter (PDEnKF) designed to assimilate out-of-range (OR) observations (soft data) explicitly: the out-of-range values are qualitative by nature (inequalities), but one can postulate a probability distribution for them and then update the ensemble members accordingly using Bayes' rule. The main idea of the PDEnKF method is that in the absence of the hard data, a virtual observation is assumed at the detection limit with a uniform OR observation likelihood in the unobservable range. The part inside from the detection limit is assumed to be Gaussian with observation error variance set equal to error variance of the hard data at the detection limit, translating the possibility that observation errors may push hard data into the unobservable range. Figure 2.2, represents the constant OR observation likelihood for an observation with an upper detection limit along with the in-range observation likelihood (Gaussian) for hard data, as assumed in Borup *et al.* (2015).

Given that the unknown observation is in the unobservable range, the assumed virtual observation at the detection limit is then only used to correct the forecast ensemble members that are inside the observable range such that the ensemble members are updated towards the detection limit. Borup *et al.* (2015) referred to this process as *partial updating* because only a part of the ensemble is updated.

The partial update of the ensemble is performed by updating the anomalies within the DEnKF. Anomalies are updated differently conditioned on the values of forecast ensemble members, that is, whether the member is inside or outside the observable range. The mean, on the other hand, is updated only when there are hard data, but is left unchanged by soft data, only algorithmically though, because the updated members will change the ensemble analysis mean when anomalies are added to the mean (final step of the DEnKF

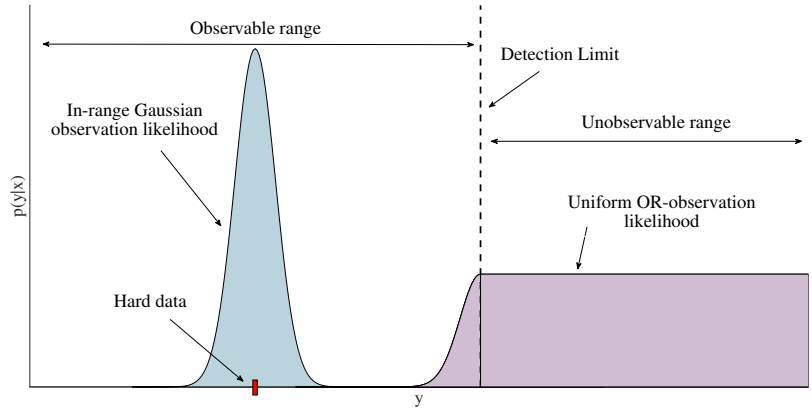


Figure 2.2: Schematic illustration of the in-range Gaussian and uniform out-of-range observation likelihood for observations with upper detection limit in the PDEnKF. Adapted from Borup *et al.* (2015)

.

algorithm). For a detailed derivation and implementation of the algorithm reader is referred to (Borup *et al.*, 2015, Section 3). The scheme has been tested using both linear and nonlinear reservoir cascade models, as used in urban water management. The authors present an important improvement in forecast accuracy, implying that soft data can contribute meaningful information to predictions.

Some issues arise from the examination of the PDEnKF algorithm: when the forecast mean is in-range but not the observations, the soft data only influence the mean through the update of anomalies instead of the update of the mean, although in principle the mean could have been updated directly. In the context of the DEnKF however, this choice is not equivalent because the implicit inflation applies to the anomalies only, not the mean. The choice of updating the anomalies by soft data rather than the mean can be seen as a "safe" choice, conform to the principles of small updates of the DEnKF and algorithmically simple to implement (M. Borup, personal communication) but possibly at the expense of reducing the impact of OR-observations. Another worry is the choice of a uniform OR-likelihood, which does not have a finite integral and is not intuitively adequate for a majority of geophysical variables: their values are getting progressively less likely as they approach the extremes. The present PhD study therefore focuses on a purely stochastic EnKF framework that does not present these difficulties.

2.3 Motivation and contribution

In DA, each observation is valuable and it is used to reduce the model uncertainty and improve forecast accuracy. The accessibility, availability and exponentially growing quantity

of observations has opened new challenges and possibility to potentially use them for inference procedures, especially in the DA field. As discussed in Section 1.2.1, observations can be divided into three categories. This thesis will focus on the study of semi-qualitative observations. Although these types of observations do not give any quantifiable data outside of its observable range, they still provide meaningful qualitative information about the observed quantity, in the shape of an inequality. In other words it provides soft data indicating that quantity is out-of-range. Geostatistical methods have proven the usefulness of soft data for the inference of static variables, but can data assimilation methods obtain the same success with dynamical systems requiring more stringent computational constrains?

As mentioned in Section 2.2.6, the PDEnKF is the only ensemble-based DA method found in the literature, which explicitly assimilates soft data. The implementation procedure of the PDEnKF is not straightforward because the mean and ensemble anomalies are updated separately. This, together with the growing availability of semi-qualitative data in various fields of climate science, motivates the work presented in this thesis. The goal of this thesis is to propose a new methodology within the family of ensemble-based DA methods, that is able to benefit from observations with a detection limit.

Similar to the work by Borup *et al.* (2015), a new ensemble-based DA methodology is proposed in this thesis to tackle the issue of observations with detection limit, but in a fully probabilistic framework. The newly proposed stochastic method is called the ensemble Kalman filter semi-qualitative (EnKF-SQ) and is developed in Paper I. The EnKF-SQ updating scheme closely follows that of the stochastic EnKF (Evensen, 2004). Apart from being a stochastic ensemble DA scheme, contrarily to the deterministic PDEnKF, the key differences between the PDEnKF and the EnKF-SQ are as follows:

- The EnKF-SQ assumes a two-piece Gaussian (Fechner, 1897; Gibbons and Mylroie, 1973) OR observation likelihood instead of the uniform OR likelihood. Because imposing a uniform density outside the observable range gives equal weight to all values until infinity, whereas extremely high values are usually less realistic in most applications, like wind speed and ice thickness among others. A two-piece Gaussian distribution is obtained by merging two opposite halves of two Gaussian probability densities (pdfs) at their common mode as follows:

$$f(x) = \begin{cases} W\exp\left(-\frac{(x-\mu)^2}{2\sigma_1^2}\right), & x \leq \mu \\ W\exp\left(-\frac{(x-\mu)^2}{2\sigma_2^2}\right), & x > \mu \end{cases} \tag{2.26}$$

  where $W = \sqrt{\frac{2}{\pi}}(\sigma_1 + \sigma_2)^{-1}$ is a normalizing constant, μ is the common mean, σ_1 and σ_2 are the standard deviations (std) of the two Gaussian pdfs. The common mean μ is located at the detection limit, as it is the last possible value the gauge

could detect with known observation uncertainty.

Figure 2.3, shows an illustration of a two-piece Gaussian OR likelihood superimposed with the uniform OR likelihood as assumed in Borup *et al.* (2015). (for further details refer to Section 2.2.2 of Shah *et al.*, 2018).

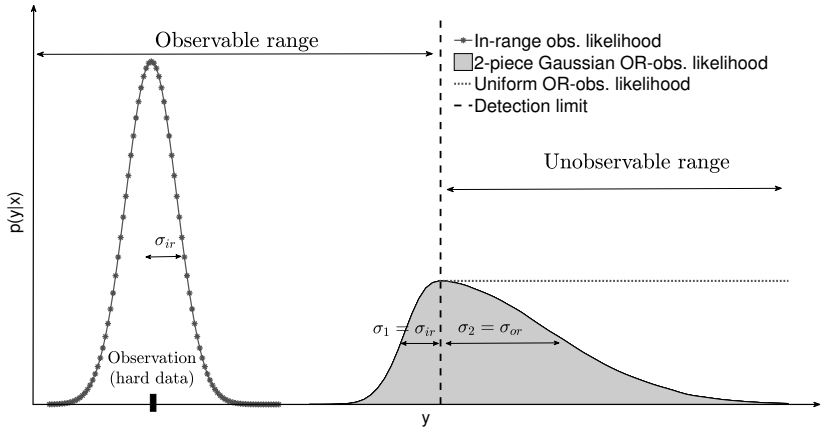


Figure 2.3: Illustration of the two-piece Gaussian OR observation likelihood, for an observation with upper detection limit. The hard data is shown by a small black rectangle and the corresponding in-range Gaussian likelihood by a solid gray line. The two-piece Gaussian likelihood is plotted in gray and a uniform OR-obs likelihood in dashed gray. σ_{ir} is the observation error standard deviation for a hard data at the detection limit and σ_{or} is the OR-observation error standard deviation.

- In the EnKF-SQ, all forecast ensemble members are updated when the observation is out-of-range, whereas for the same situation the PDEnKF only partially updates the members inside the observable range. The two EnKF-SQ update equations for the forecast ensemble members in the absence of hard data are proposed on the basis of the Bayesian update of the Gaussian prior with a two-piece Gaussian OR likelihood, although they do not coincide strictly with the Bayesian posterior (Section 2.2.3; Shah *et al.*, 2018). The update of the forecast ensemble members for soft data is divided into two cases, depending whether the observed ensemble member, $\mathbf{H}\mathbf{x}_i^f$, is inside or outside the observable range. Members inside (outside) the observable range should be updated linearly using the EnKF update equations with the observation uncertainty σ_{ir} (σ_{or}). In essence, the EnKF-SQ uses two different Kalman gains depending on the value of the observed ensemble member. For a scalar case, the in-range Kalman gain becomes $K_{ir} = \sigma_b^2\left(\sigma_b^2 + \sigma_{ir}^2\right)^{-1}$. If the member is outside the observable range then the Kalman gain is calculated with out-of-range observation error variance σ_{or} i.e., $K_{or} = \sigma_b^2\left(\sigma_b^2 + \sigma_{or}^2\right)^{-1}$. Here, the forecast error variance is denoted by σ_b^2. This way, the algorithm updates the entire forecast ensemble instead of partial updates and does not split the update of the mean and anomalies, thus

avoiding any ambiguity.

Paper I presents the detailed derivation of the EnKF-SQ along with results of numerical experiments performed with linear and nonlinear toy models under twin-experiments. The performance of the EnKF-SQ is compared with the PDEnKF. The numerical results show that assimilating qualitative observations using the proposed scheme improves the overall forecast mean. The results also indicate that with the same configuration of the ensemble size and other parameters, the EnKF-SQ remains robust and outperforms the PDEnKF. The proposed scheme specifically improves the forecast accuracy in the vicinity of the detection limit in the absence of the hard data by adding value from the soft data. This feature of the scheme makes it attractive compare to just ignoring the soft data.

In order to assess the viability and performance of the newly proposed EnKF-SQ in a high-dimensional complex system, it is implemented using the state-of-the-art coupled Arctic ocean sea ice model TOPAZ4 configuration (Sakov *et al.*, 2012) assimilating sea ice thickness with a detection limit (similar to the SMOS product). The details of the experimental configuration and numerical results are presented in Paper II. Various experiments are performed and the results suggest that the EnKF-SQ clearly makes a valuable approach for assimilating semi-qualitative observations into high-dimensional nonlinear systems.

The EnKF-SQ algorithm explicitly depends on a climatology of the OR values of the observed quantity to estimate the most important parameter, and the only new free parameter of the method, the OR observation error variance of the two-piece Gaussian likelihood. The dependency of the OR observation error variance estimate on the climatology, makes the EnKF-SQ sensitive to imprecise climatology values. To compensate for an imprecise OR-observation error variance specification, an adaptive spatially and temporally varying OR-observation error variance correction scheme is proposed within the framework of the EnKF-SQ. Paper III presents the formulation of this correction strategy along with results of numerical experiments using a nonlinear Lorenz'96 toy model. The adaptive scheme shows improvements of the overall forecast accuracy when compared to the fixed OR-observation error variance both in biased and unbiased twin experiments, even for extreme climatological biases.

The PhD candidate is the main author of all publications, written under the supervision and in collaboration with the PhD supervisors and other colleagues at NERSC.

Chapter 3

Summary of papers and outlook

With this chapter, we end the introduction of the thesis by briefly summarizing each of the articles included in it and discussing further work and potential improvements.

3.1 Paper I summary

Assimilation of semi-qualitative observations with a stochastic ensemble Kalman filter

Shah, A., El Gharamti, M., and Bertino, L.

Published in *Quarterly Journal of the Royal Meteorological Society*

The first article introduces an ensemble-based data assimilation method that addresses the problem of observations with a detection limit. Most data assimilation methods discards the out-of-range (OR) values, treating them as not a number, with loss of possibly useful qualitative information (soft data). Inspired by the study of Borup *et al.* (2015), we propose a fully probabilistic ensemble-based data assimilation method namely ensemble Kalman filter - semi qualitative (EnKF-SQ), which explicitly assimilates soft data.

Whenever the available observation is OR, a virtual observation at the detection limit is created assuming a two-piece Gaussian observation likelihood around it. The mode of the two-piece Gaussian likelihood is imposed at the detection limit; the variance inside the observable range is equal to the observation error variance (σ_{ir}^2) of hard data at the detection limit, whereas the variance in the unobservable range (out-of-range error variance σ_{or}^2) is defined with the help of a climatology.

A Bayesian approach is applied to examine the posterior distribution of the model state variables. A Gaussian prior distribution is updated with an assumed asymmetric two-piece Gaussian OR observational likelihood to obtain the posterior distribution. The EnKF-SQ al-

gorithm is developed on the basis of the Bayesian update and closely follows the stochastic EnKF (Burgers *et al.*, 1998). The central idea of the EnKF-SQ algorithm is the computation of a separate Kalman gain matrix **K** for updating each ensemble member rather than using a global one. This is also the main algorithmic difference compared to the stochastic EnKF update scheme. In order to evaluate the EnKF-SQ posterior distribution against the true Bayesian posterior, a one-dimensional example with a Gaussian prior and a two-piece OR Gaussian likelihood is used.

Finally, the EnKF-SQ is tested in a twin-experiment framework using a linear subsurface flow-transport model (obeying Darcy's law) and the nonlinear Lorenz'96 model assimilating observations with an upper detection limit. The root mean square errors and average ensemble spread of the forecast estimates are used to evaluate the performance of the scheme. Sensitivity experiments with varying ensemble size and detection limit (changing the number of observations falling out-of-range) are conducted to check the robustness of the scheme. In addition, sensitivity experiments with varying σ^2_{or} are also conducted to study its effect on the performance of the EnKF-SQ and specifically on the higher-order moments of the posterior distribution.

Our numerical results show that assimilating qualitative observations using the EnKF-SQ improves the overall forecast skill.

3.2 Paper II summary

Assimilation of semi-qualitative sea ice thickness data with the EnKF-SQ

Shah, A., Bertino, L., Counillon, F., El Gharamti, M., and Xie, J.

Submitted in *Tellus A: Dynamic Meteorology and Oceanography*

Paper II is a follow up study of Paper I where the EnKF-SQ is applied to a realistic state-of-the-art coupled ice-ocean model of the Arctic, the TOPAZ4 configuration, in a twin-experiment framework. Synthetic thin sea ice thickness (SIT) data with an upper detection limit of 1.0 m are assimilated, mimicking SMOS retrieved SIT observations. The goal of the study is to check the feasibility of the implementation of the EnKF-SQ in such a high-dimensional complex system. The method is shown to add value to range-limited thin ice thickness measurements, as obtained from passive microwave remote sensing, with respect to more trivial solutions like neglecting the out-of-range values or assimilating climatology values instead.

In order to generate the synthetic SIT data mimicking SMOS SIT, a reference *truth* run is produced by integrating the coupled ocean sea ice model for two years using unperturbed

atmospheric reanalysis forcing. Synthetic SIT observations are then created by perturbing the truth with white Gaussian noise. Further, a synthetic OR SIT climatology is computed, to be assimilated instead of soft data using the standard EnKF, by taking the two-year time average of all *true SIT* above the detection limit in each grid cell.

The EnKF-SQ is tested for the first time with a high-dimensional system and the implementation was intentionally simplified: the local state vector therefore consists of only two sea-ice variables: SIT and sea-ice concentration, constituting a case of a weakly coupled but multivariate assimilation. The model error increasing the model spread is introduced via perturbing the atmospheric forcing fields. The synthetic SIT observations are assimilated with a detection limit of 1.0 m every week and a local analysis is performed in which the two variables at each grid cell are updated using only the nearest observation. Different single-cycle sensitivity experiments are performed with varying ensemble size and OR observation error variance of the two-piece Gaussian OR observation likelihood (the only free parameter of the EnKF-SQ), to settle their values for the long assimilation experiment. The results from the sensitivity experiments confirm the findings from Paper I.

A 5-months DA experiment was performed during the winter 2014-2015 using 99 ensemble members in the EnKF-SQ and other EnKF benchmarks. Different assimilation experiments are conducted to assess the performance of the EnKF-SQ against other EnKF configurations assimilating (1) only thin ice; (2) both thin and thick ice; and (3) climatology. The study shows that assimilating soft data improves the SIT forecast accuracy compared to ignoring them by approximately 8%, particularly where sea ice approaches the detection limit. Such a difference can be important in light of the performance of the TOPAZ4 operational system. The performance exhibited by assimilating a reasonably accurate climatology was similar to the EnKF-SQ. Assessing the bias of the analysis did not reveal the introduction of any significant biases by the EnKF-SQ.

3.3 Paper III summary

An adaptive correction algorithm for the out-of-range observation error variance of the EnKF-SQ

Shah, A., El Gharamti, M., Bertino, L., and Counillon, F.
To be submitted

Paper III focuses on an algorithmic modification to the proposed EnKF-SQ scheme. The formulation of the EnKF-SQ depends on a likelihood probability distribution for out-of-range observations, parameterised by an out-of-range observation error variance σ^2_{or}, usually computed from a climatology. The climatology however can be biased, which may affect

the performance of the EnKF-SQ. This article presents a temporally and spatially adaptive σ^2_{or} correction strategy for the EnKF-SQ to alleviate the effect of a biased climatology.

Inspired by the adaptive inflation schemes of Anderson (2007) and El Gharamti (2018), a Bayesian approach is used, where σ^2_{or} is considered a random variable and sequentially estimated at every assimilation cycle using the data. Since, σ^2_{or} is a positive quantity, an inverse gamma prior distribution is assumed, which is updated given the data likelihood (coming from the pdf of innovations) to obtain a posterior pdf. The posterior is then maximized to find the mode of the distribution, which is then selected as the new estimate of σ^2_{or} and then reused in the prior for next assimilation cycle. Independent values of σ^2_{or} are computed independently for all state variables and therefore the correction strategy is adaptive in time and space.

The newly proposed algorithm is called EnKF-SQ-Adap to distinguish it from the EnKF-SQ, which uses a fixed σ^2_{or}. The algorithmic implementation of the EnKF-SQ-Adap adds the new σ^2_{or} correction strategy between the forecast and update step of the EnKF-SQ. Apart from this new modification, there is no change in the EnKF-SQ code.

The EnKF-SQ-Adap is evaluated in both biased and unbiased twin-experiment setups using the Lorenz'96 model in cases of biased climatology. All possible combinations of model biases and climatology biases are thus covered. The performance is compared to the pre-existing EnKF-SQ and the same three EnKF benchmarks in Paper II: (1) only assimilating hard data (EnKF-IG, Ignore), (2) no detection limit (EnKF-ALL) and (3) assimilating climatology instead of soft data (EnKF-CLIM).

The numerical results compared to the EnKF-SQ, EnKF-CLIM and EnKF-IG suggest that the EnKF-SQ-Adap improves the overall forecast accuracy in biased as well as unbiased conditions for all choices of climatology. Further, the EnKF-SQ-Adap was found robust to DA forward model bias and suffered less filter divergence than the EnKF-SQ and EnKF-CLIM, making the EnKF-SQ-Adap a preferable choice over its predecessor, the EnKF-SQ and also the simple EnKF-CLIM. Despite the good performance of the EnKF-SQ-Adap, the scheme still performs poorly when the DA forecast model is extremely biased because the adaptive scheme carries the model bias over to σ^2_{or} estimate. Such cases should be better treated by methods explicitly targeted at model bias estimation.

3.4 Further work and perspectives

This thesis has shown that the newly proposed EnKF-SQ and its variant the adaptive EnKF-SQ-Adap makes better use of semi-qualitative data rather than ignoring the data above the detection limit, at a reasonable computing cost. It is also shown that the EnKF-SQ is able to outperform the assimilation of climatology for values above the detection limit under cer-

tain conditions and perform equally good otherwise. The limitations of the methods were emphasized along with the potential benefits. The methods have been tested under biased and unbiased twin-experiments using a range of linear, nonlinear, low to high dimensional complex dynamical systems. Numerical results from all these experiments clearly suggest the benefits of assimilating soft data with the EnKF-SQ DA scheme. Further steps include, ordered from short-term to long-term perspectives:

- In paper II, the EnKF-SQ was implemented with a coupled ocean sea ice model, where for the sake of implementation simplicity the state vector only included two variables. This implementation can be scaled up to include all state variables from the ocean and sea ice but this will also necessitate optimizing the computations for larger state vector.

- Following the previous point on increasing the level of complexity, the EnKF-SQ should be included into an existing data assimilation package like (among others) the NERSC EnKF, DART from NCAR, the parallel data assimilation framework (PDAF) from the Alfred Wegener Institute or the EnKF-C from the Bureau of Meteorology, Australia.

- The thesis presents a case for assimilating synthetic SIT data with the EnKF-SQ. The next step is to assimilate real SMOS data which comes at a different resolution than the one used in the synthetic case. Besides, by using a lower threshold rather than a higher threshold, the complementary CryoSAT2 thick SIT can be included as well. Assimilating both thin and thick SIT separately would help respecting the sampling and observation error characteristics of both satellites in contrast to the assimilation of a merged spatio-temporally interpolated product (Ricker *et al.*, 2017).

- Implementing the EnKF-SQ with other realistic applications in Earth system sciences, like observations of chlorophyll with a lower detection limit in ocean biogeochemistry, soil moisture from SMOS with a higher detection limit among others. This should help understand how well the scheme can tackle various kinds of observation detection limits but also allow for further developments of the scheme that are unexplored here.

- In addition to semi-qualitative data, the EnKF-SQ can also be expanded to the assimilation of fully qualitative data. For example the ice types mentioned earlier, as well as other binary indices that are common in Earth system sciences (is a contaminant present or not, is permafrost present or not). Such data can often be related to specific thresholds of a hidden quantitative state variable: for example the sea-ice types can be considered a classification of sea-ice age: young ice between zero and 30 days, first-year ice up to 365 days and multi-year ice above. Such multiple detection limits

can then be used to formulate an OR observation likelihood. Now if the data suggests that the quantity is on either side of the detection limit, the forecast ensemble members falling on the other side of the observation shall be corrected towards the detection limit in the direction of the qualitative observation. The stochastic framework used in this thesis can in principle be extended to tackle this problem, but the algorithmic complexity - the number of particular cases to be taken into account - will increase considerably.

- The EnKF-SQ uses a two-piece Gaussian OR observation likelihood for observation out-of-range. This is not necessarily a best choice for every application. Therefore, replacing the two-piece Gaussian by other distributions depending on the application will be interesting to study.

Bibliography

Aaboe S, Breivik LA, Sørensen A, Eastwood S, Lavergne T. 2018. Global sea ice edge and type product user's manual, product OSI-402-c & OSI-403-c. Technical report, Norwegian Meteorological Institute, URL http://osisaf.met.no/docs/osisaf_cdop3_ss2_pum_sea-ice-edge-type_v2p3.pdf.

Aanonsen SI, Nævdal G, Oliver DS, Reynolds AC, Vallès B. 2009. The ensemble Kalman filter in reservoir engineering - a review. *SPE Journal* **14**(03): 393–412, doi:10.2118/117274-PA, URL https://doi.org/10.2118/117274-PA.

Anderson JL. 2001. An ensemble adjustment Kalman filter for data assimilation. *Monthly weather review* **129**(12): 2884–2903.

Anderson JL. 2007. An adaptive covariance inflation error correction algorithm for ensemble filters. *Tellus A: Dynamic Meteorology and Oceanography* **59**(2): 210–224, doi:10.1111/j.1600-0870.2006.00216.x, URL https://doi.org/10.1111/j.1600-0870.2006.00216.x.

Asch M, Bocquet M, Nodet M. 2016. *Data Assimilation: Methods, Algorithms, and Applications*. Fundamentals of Algorithms, SIAM, Philadelphia, ISBN 978-1-611974-53-9.

Bauer P, Thorpe A, Brunet G. 2015. The quiet revolution of numerical weather prediction. *Nature* **525**(7567): 47.

Birkett CM. 1998. Contribution of the TOPEX NASA radar altimeter to the global monitoring of large rivers and wetlands. *Water Resources Research* **34**(5): 1223–1239, doi:10.1029/98WR00124, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/98WR00124.

Bishop CH, Etherton BJ, Majumdar SJ. 2001. Adaptive sampling with the ensemble transform Kalman filter. Part I: theoretical aspects. *Monthly Weather Review* **129**(3): 420–436, doi:10.1175/1520-0493(2001)129<0420:ASWTET>2.0.CO;2, URL https://doi.org/10.1175/1520-0493(2001)129<0420:ASWTET>2.0.CO;2.

Bocquet M, Pires CA, Wu L. 2010. Beyond Gaussian statistical modeling in geophysical data assimilation. *Monthly Weather Review* **138**(8): 2997–3023.

Bocquet M, Sakov P. 2014. An iterative ensemble Kalman smoother. *Quarterly Journal of the Royal Meteorological Society* **140**(682): 1521–1535, doi:10.1002/qj.2236, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.2236.

Borup M, Grum M, Madsen H, Mikkelsen PS. 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, URL http://dx.doi.org/10.1007/s00477-014-0908-1.

Buehner M, Houtekamer P, Charette C, Mitchell HL, He B. 2010a. Intercomparison of variational data assimilation and the ensemble Kalman filter for global deterministic NWP. Part II: one-month experiments with real observations. *Monthly Weather Review* **138**(5): 1567–1586.

Buehner M, Houtekamer PL, Charette C, Mitchell HL, He B. 2010b. Intercomparison of variational data assimilation and the ensemble Kalman filter for global deterministic NWP. Part I: description and single-observation experiments. *Monthly Weather Review* **138**(5): 1550–1566, doi:10.1175/2009MWR3157.1, URL https://doi.org/10.1175/2009MWR3157.1.

Burgers G, Jan van Leeuwen P, Evensen G. 1998. Analysis scheme in the ensemble Kalman filter. *Monthly weather review* **126**(6): 1719–1724.

Carrassi A, Bocquet M, Bertino L, Evensen G. 2018. Data assimilation in the geosciences: An overview of methods, issues, and perspectives. *Wiley Interdiscip. Rev. Clim. Chang.* **9**(5): e535, doi:10.1002/wcc.535, URL http://doi.wiley.com/10.1002/wcc.535.

Chiles JP, Delfiner P. 2012. *Geostatistics: modeling spatial uncertainty (2nd ed)*. Wiley: New York, ISBN 0471083151 9780471083153.

Counillon F, Keenlyside N, Bethke I, Wang Y, Billeau S, Shen ML, Bentsen M. 2016. Flow-dependent assimilation of sea surface temperature in isopycnal coordinates with the norwegian climate prediction model. *Tellus A: Dynamic Meteorology and Oceanography* **68**(1): 32 437, doi:10.3402/tellusa.v68.32437, URL https://doi.org/10.3402/tellusa.v68.32437.

Dee DP, Uppala SM, Simmons AJ, Berrisford P, Poli P, Kobayashi S, Andrae U, Balmaseda MA, Balsamo G, Bauer P, Bechtold P, Beljaars ACM, van de Berg L, Bidlot J, Bormann N, Delsol C, Dragani R, Fuentes M, Geer AJ, Haimberger L, Healy SB, Hersbach H, Hólm EV, Isaksen L, Kållberg P, Köhler M, Matricardi M, McNally AP, Monge-Sanz BM, Morcrette JJ, Park BK, Peubey C, de Rosnay P, Tavolato C, Thépaut JN, Vitart F. 2011. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the royal meteorological society* **137**(656): 553–597.

El Gharamti M. 2018. Enhanced adaptive inflation algorithm for ensemble filters. *Monthly Weather Review* **146**(2): 623–640, doi:10.1175/MWR-D-17-0187.1, URL https://doi.org/10.1175/MWR-D-17-0187.1.

Emery X, Arroyo D, Peláez M. 2014. Simulating Large Gaussian Random Vectors Subject to Inequality Constraints by Gibbs Sampling. *Math. Geol. Geosci.* **46**: 265–283, doi:10.1007/s11004-013-9495-9.

Emery X, Robles LN. 2008. Simulation of mineral grades with hard and soft conditioning data: application to a porphyry copper deposit. *Computational Geosciences* **13**(1): 79, doi:10.1007/s10596-008-9106-x, URL https://doi.org/10.1007/s10596-008-9106-x.

Evensen G. 1992. Using the Extended {K}alman Filter with a Multilayer Quasi-Geostrophic Ocean Model. *J. Geophys. Res.* **97**(C11): 17 905–17 924.

Evensen G. 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte-Carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans* **99**(C5): 10 143–10 162.

Evensen G. 2003. The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean dynamics* **53**(4): 343–367.

Evensen G. 2004. Sampling strategies and square root analysis schemes for the EnKF. *Ocean Dynamics* **54**(6): 539–560, doi:10.1007/s10236-004-0099-2, URL https://doi.org/10.1007/s10236-004-0099-2.

Fechner G. 1897. *Kollektivmasslehre (ed. g. f. lipps)*. Auftrage der königl. Sächsichen Gesellschaft der Wissenschaften (in German).

Gharamti M, Ait-El-Fquih B, Hoteit I. 2015. An iterative ensemble Kalman filter with one-step-ahead smoothing for state-parameters estimation of contaminant transport models. *Journal of Hydrology* **527**: 442–457.

Gharamti M, Valstar J, Hoteit I. 2014. An adaptive hybrid EnKF-OI scheme for efficient state-parameter estimation of reactive contaminant transport models. *Advances in water resources* **71**: 1–15.

Ghil M, Malanotte-Rizzoli P. 1991. Data assimilation in meteorology and oceanography. *Advances in geophysics* **33**: 141–266.

Gibbons J, Mylroie S. 1973. Estimation of impurity profiles in ion implanted amorphous targets using joined half Gaussian distributions. *Applied Physics Letters* **22**(11): 568–569, doi:10.1063/1.1654511, URL http://dx.doi.org/10.1063/1.1654511.

Hamill TM, Snyder C. 2000. A Hybrid Ensemble Kalman Filter–3D Variational Analysis Scheme. *Monthly Weather Review* **128**(8): 2905–2919, doi:10.1175/1520-0493(2000)128<2905:AHEKFV>2.0.CO;2, URL https://doi.org/10.1175/1520-0493(2000)128<2905:AHEKFV>2.0.CO;2.

Hornung RW, Reed LD. 1990. Estimation of average concentration in the presence of nondetectable values. *Applied occupational and environmental hygiene* **5**(1): 46–51.

Houtekamer PL, Mitchell HL. 1998. Data assimilation using an ensemble Kalman filter technique. *Monthly Weather Review* **126**(3): 796–811, doi:10.1175/1520-0493(1998)126<0796:DAUAEK>2.0.CO;2, URL https://doi.org/10.1175/1520-0493(1998)126<0796:DAUAEK>2.0.CO;2.

Ivanova N, Pedersen LT, Tonboe RT, Kern S, Heygster G, Lavergne T, Sørensen A, Saldo R, Dybkjær G, Brucker L, Shokr M. 2015. Inter-comparison and evaluation of sea ice algorithms: towards further identification of challenges and optimal approach using passive microwave observations. *Cryosph.* **9**(5): 1797–1817, doi:10.5194/tc-9-1797-2015, URL http://www.the-cryosphere.net/9/1797/2015/.

Jazwinski A. 1970. *Stochastic processes and filtering theory*, vol. 63. Academic Press, first edn.

Journel AG. 1986. Constrained interpolation and qualitative information—the soft kriging approach. *Mathematical Geology* **18**(3): 269–286, doi:10.1007/BF00898032, URL https://doi.org/10.1007/BF00898032.

Journel AG, Huijbregts CJ. 1978. *Mining Geostatistics*. Academic Press.

Kaleschke L, Tian-Kunze X, Maaß N, Mäkynen M, Drusch M. 2012. Sea ice thickness retrieval from SMOS brightness temperatures during the Arctic freeze-up period. *Geophysical Research Letters* **39**(5): n/a–n/a, doi:10.1029/2012GL050916, URL http://dx.doi.org/10.1029/2012GL050916. L05501.

Kalman RE, *et al.* 1960. A new approach to linear filtering and prediction problems. *Journal of basic Engineering* **82**(1): 35–45.

Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, Iredell M, Saha S, White G, Woollen J, Zhu Y, Chelliah M, Ebisuzaki W, Higgins W, Janowiak J, Mo KC, Ropelewski C, Wang J, Leetmaa A, Reynolds R, Jenne R, Joseph D. 1996. The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society* **77**(3): 437–472, doi:10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2, URL https://doi.org/10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2.

Laloyaux P, de Boisseson E, Balmaseda M, Bidlot JR, Broennimann S, Buizza R, Dalhgren P, Dee D, Haimberger L, Hersbach H, Kosaka Y, Martin M, Poli P, Rayner N, Rustemeier E, Schepers D. 2018. CERA-20C: A Coupled Reanalysis of the Twentieth Century. *Journal of Advances in Modeling Earth Systems* **10**(5): 1172–1195, doi:10.1029/2018MS001273, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018MS001273.

Laxon SW, Giles KA, Ridout AL, Wingham DJ, Willatt R, Cullen R, Kwok R, Schweiger A, Zhang J, Haas C, Hendricks S, Krishfield R, Kurtz N, Farrell S, Davidson M. 2013. Cryosat-2 estimates of arctic sea ice thickness and volume. *Geophysical Research Letters* **40**(4): 732–737, doi:10.1002/grl.50193, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/grl.50193.

Le Dimet FX, Talagrand O. 1986. Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. *Tellus A: Dynamic Meteorology and Oceanography* **38**(2): 97–110.

Le Gland F, Monbet V, Tran VD. 2009. Large sample asymptotics for the ensemble Kalman filter. Research Report RR-7014, INRIA, URL https://hal.inria.fr/inria-00409060.

Lewis JM, Derber JC. 1985. The use of adjoint equations to solve a variational adjustment problem with advective constraints. *Tellus A* **37**(4): 309–322.

Liu C, Xiao Q, Wang B. 2008. An ensemble-based four-dimensional variational data assimilation scheme. Part I: technical formulation and preliminary test. *Monthly Weather Review* **136**(9): 3363–3373, doi:10.1175/2008MWR2312.1, URL https://doi.org/10.1175/2008MWR2312.1.

Lorenc AC. 1986. Analysis methods for numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society* **112**(474): 1177–1194, doi:10.1002/qj.49711247414, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.49711247414.

Macelloni G, Brogioni M, Leduc-Leballeur M, Montomoli F, Bartsch A, Mialon A, Ritz C, Closa Soteras J, Stammer D, Picard G, De Carolis G, Boutin J, Johnson J, Nicholls K, Jezek K, Rautiainen K, Kaleschke L, Bertino L, Tsang L, van den Broeke M, Skou N, Tietsche S. 2018. Cryorad: A low frequency wideband radiometer mission for the study of the cryosphere. In: *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, pp. 1998–2000, doi:10.1109/IGARSS.2018.8519172.

Marcotte D, Allard D. 2018. Gibbs sampling on large lattice with GMRF. *Comput. Geosci.* **111**(June 2017): 190–199, doi:10.1016/j.cageo.2017.11.012, URL https://doi.org/10.1016/j.cageo.2017.11.012.

Oliver DS, Chen Y. 2011. Recent progress on reservoir history matching: a review. *Computational Geosciences* **15**(1): 185–221, doi:10.1007/s10596-010-9194-2, URL https://doi.org/10.1007/s10596-010-9194-2.

Pasetto D, Finger F, Rinaldo A, Bertuzzo E. 2017. Real-time projections of cholera outbreaks through data assimilation and rainfall forecasting. *Advances in water resources* **108**: 345–356.

Patilea C, Heygster G, Huntemann M, Spreen G. 2019. Combined SMAP–SMOS thin sea ice thickness retrieval. *The Cryosphere* **13**(2): 675–691, doi:10.5194/tc-13-675-2019, URL https://www.the-cryosphere.net/13/675/2019/.

Raanes PN, Bocquet M, Carrassi A. 2019. Adaptive covariance inflation in the ensemble Kalman filter by Gaussian scale mixtures. *Quarterly Journal of the Royal Meteorological Society* **145**(718): 53–75, doi:10.1002/qj.3386.

Reale A, Tilley F, Ferguson M, Allegrino A. 2008. NOAA operational sounding products for advanced TOVS. *International Journal of Remote Sensing* **29**(16): 4615–4651, doi:10.1080/01431160802020502, URL https://doi.org/10.1080/01431160802020502.

Rhodes C, Hollingsworth T. 2009. Variational data assimilation with epidemic models. *Journal of Theoretical Biology* **258**(4): 591 – 602, doi:https://doi.org/10.1016/j.jtbi.2009.02.017, URL http://www.sciencedirect.com/science/article/pii/S0022519309000794.

Ricker R, Hendricks S, Kaleschke L, Tian-Kunze X, King J, Haas C. 2017. A weekly Arctic sea-ice thickness data record from merged CryoSat-2 and SMOS satellite data. *The Cryosphere* **11**(4): 1607–1623, doi:10.5194/tc-11-1607-2017, URL https://www.the-cryosphere.net/11/1607/2017/.

Sakov P, Counillon F, Bertino L, Lisæter K, Oke P, Korablev A. 2012. TOPAZ4: an ocean-sea ice data assimilation system for the north Atlantic and Arctic. *Ocean Science* **8**(4): 633.

Sakov P, Oke PR. 2008. A deterministic formulation of the ensemble Kalman filter: an alternative to ensemble square root filters. *Tellus A* **60**(2): 361–371.

Shah A, Bertino L, Counillon F, Gharamti ME, Xie J. 2019. Assimilation of semi-qualitative sea ice thickness data with the EnKF-SQ. *arXiv preprint arXiv:1904.12590* .

Shah A, Gharamti ME, Bertino L. 2018. Assimilation of semi-qualitative observations with a stochastic ensemble kalman filter. *Quarterly Journal of the Royal Meteorological Society* **0**(0), doi:10.1002/qj.3381, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3381.

Shapiro M, Shukla J, Hoskins B, Church J, Trenberth K, Béland M, Brasseur G, Wallace M, McBean G, Caughey J, *et al.* 2009. The socioeconomic and environmental benefits of a weather, climate and earth-system prediction initiative for the 21st century. *Bulletin of the American Meteorological Society* .

Whitaker JS, Hamill TM. 2002. Ensemble data assimilation without perturbed observations. *Monthly Weather Review* **130**(7): 1913–1924.

Wikle CK, Berliner LM. 2007. A Bayesian tutorial for data assimilation. *Physica D: Nonlinear Phenomena* **230**(1): 1 – 16, doi:https://doi.org/10.1016/j.physd.2006.09.017, URL http://www.sciencedirect.com/science/article/pii/S016727890600354X.

Williamson RA, Hertzfeld HR, Cordes J. 2002. The socio-economic value of improved weather and climate information. *Space Policy Institute, Washington, DC. Available at https://esto.nasa.gov/files/2001/Economic%20Study/Socio-EconomicBenefitsFinalREPORT1.pdf [Verified 22 June 2019]* .

Xie J, Bertino L, Counillon F, Lisæter KA, Sakov P. 2017. Quality assessment of the TOPAZ4 reanalysis in the Arctic over the period 1991–2013. *Ocean Science* **13**(1): 123–144, doi:10.5194/os-13-123-2017, URL https://www.ocean-sci.net/13/123/2017/.

Part II

Research Articles

Paper I

Assimilation of semi-qualitative observations with a stochastic ensemble Kalman filter

Shah, A., El Gharamti, M., and Bertino, L.
Quarterly Journal of the Royal Meteorological Society, published

Received: 20 March 2018 | Revised: 21 June 2018 | Accepted: 12 July 2018

DOI: 10.1002/qj.3381

RESEARCH ARTICLE

Quarterly Journal of the Royal Meteorological Society RMetS

Assimilation of semi-qualitative observations with a stochastic ensemble Kalman filter

Abhishek Shah[1] | Mohamad El Gharamti[1,2] | Laurent Bertino[1]

[1]Nansen Environmental and Remote Sensing Center, Bergen, Norway
[2]Computational and Information Systems Laboratory, Data Assimilation Research Section National Center for Atmospheric Research, Boulder, CO, USA

Correspondence
Abhishek Shah, Nansen Environmental and Remote Sensing Center, Thormøhlensgate 47, 5006 Bergen, Norway.
Email: abhishek.shah@nersc.no

The ensemble Kalman filter assumes observations to be Gaussian random variables with a pre-specified mean and variance. In practice, observations may also have detection limits, for instance when a gauge has a minimum or maximum value. In such cases, most data assimilation schemes discard out-of-range values, treating them as "not a number," with the loss of possibly useful qualitative information.
The current work focuses on the development of a data assimilation scheme that tackles observations with a detection limit. We present the Ensemble Kalman Filter Semi-Qualitative (EnKF-SQ) and test its performance against the Partial Deterministic Ensemble Kalman Filter (PDEnKF) of Borup *et al.* Both are designed to assimilate out-of-range observations explicitly: the out-of-range values are qualitative by nature (inequalities), but one can postulate a probability distribution for them and then update the ensemble members accordingly. The EnKF-SQ is tested within the framework of twin experiments, using both linear and nonlinear toy models. Different sensitivity experiments are conducted to assess the influence of the ensemble size, observation detection limit and number of observations on the performance of the filter. Our numerical results show that assimilating qualitative observations using the proposed scheme improves the overall forecast mean, making it viable for testing on more realistic applications such as sea-ice models.

KEYWORDS

data assimilation, detection limit, ensemble Kalman filter, semi-qualitative information, out-of-range observations

1 | INTRODUCTION

Data Assimilation (DA) is an approach through which available observations, along with prior knowledge (model state), are used to obtain an estimate of the true state of a process (Ghil and Malanotte-Rizzoli, 1991; Daley, 1993; Talagrand, 1997; Kalnay, 2003). Each observation is used to reduce model uncertainty and improve forecast accuracy. In practice, many observations are only available in a limited interval of the actual variation of the observed quantity, that is, observations with a detection limit. For instance, some observations with a higher detection limit are Soil Moisture and Ocean Salinity (SMOS) satellite estimates of the sea-ice thickness (Kaleschke *et al.*, 2010; 2012) and ocean wind observations from scatterometers at hurricane wind speeds (Reul *et al.*, 2012). SMOS can give quantitative thickness data only up to 50 cm over first-year level ice for the Arctic, because the signal penetration is limited by the wavelength. In reality, the sea ice can grow up to a few metres. Conversely, observations with lower detection limit also exist. Examples are contaminant concentrations in environmental and health fields (Hornung and Reed, 1990) and river water level measurements obtained from satellite radar altimetry. On top of the detection limits, some measurements are Boolean in nature: for example, whether permafrost exists or not (Li and Cheng, 1999), or whether or not there is overflow at a weir in urban hydrology (Thorndahl *et al.*, 2008). Although these types of observation do not provide quantifiable data above or below the detection limit, they do give qualitative information about the observed variable. Therefore, this type of observation should be exploited as a means to improve the model forecast.

All deterministic and stochastic ensemble-based filtering schemes (Burgers *et al.*, 1998; Anderson, 2001; Tippett *et al.*, 2003; Sakov and Oke, 2008) assimilate actual observations

(hard data), but do not consider qualitative information (soft data) available from out-of-range observations (OR observations). Whereas the geostatistical techniques are well established for variables without dynamical evolution (Chiles and Delfiner, 1999; Emery and Robles, 2008), only one study (Borup *et al.*, 2015), to the best of our knowledge, has dealt with the issue of OR observations in an ensemble-based data assimilation framework. The issue has been addressed in variational methods (see Bocquet *et al.*, 2010, section 2c and references therein).

Borup *et al.* (2015) proposed the Partial Deterministic Ensemble Kalman Filter (PDEnKF) to assimilate observations with a detection limit. The main idea of the PDEnKF is to assume a virtual observation at the detection limit in the absence of hard data and defining a constant OR observation likelihood in the unobservable region from the detection limit. The virtual observation is then used to update the anomalies within the framework of the Deterministic Ensemble Kalman Filter (DEnKF: Sakov and Oke, 2008). Anomalies are updated differently conditioned on the values of forecast ensemble members, that is, whether the member is inside or outside the observable range. The mean, on the other hand, is updated only when there are hard data, or else there is no update. In practice, virtual OR observations are used only to update the ensemble members that are within the observable range. The scheme has been tested using both linear and nonlinear reservoir cascade models. The authors present an important improvement in forecast accuracy, implying that soft data can contribute meaningful information to predictions.

In light of this background, a new DA algorithm, referred to as the Ensemble Kalman Filter Semi-Qualitative (EnKF-SQ), is developed here and is designed to assimilate OR observations explicitly. The EnKF-SQ assumes a virtual observation at the detection limit in the absence of hard data, with an asymmetric two-piece Gaussian observational likelihood on either side of the detection limit. In the EnKF-SQ, the forecast ensemble members are updated by the observations, which are perturbed using a two-piece Gaussian OR observation likelihood following the stochastic ensemble Kalman filter (EnKF) update (Evensen, 2003). A detailed derivation of the EnKF-SQ is discussed in section 2.2, followed by an algorithmic implementation. To test the performance of the EnKF-SQ, we apply it to two different linear and nonlinear toy models. The experimental setup and results are presented in section 3. A summary of the numerical results is followed by a general discussion that concludes the article in section 4.

2 | METHODOLOGY AND ALGORITHM

In this section, a brief background on the stochastic EnKF is given. The new EnKF-SQ is also derived and presented in detail.

2.1 | Background

The Kalman filter (KF: Kalman, 1960) is a sequential filtering technique, in which the model is integrated forward in time and, when they become available, observations are used to update the model state and its associated uncertainty. The KF is a recursive Bayesian estimation method, which is optimal for Gaussian and linear models (Kalman, 1960; Gharamti *et al.*, 2012, and references therein). The KF operates sequentially in time, following time update (forecast) and measurement update (analysis) steps. The EnKF, a variant of the KF, utilizes an ensemble of model states $[\mathbf{x}_1, \mathbf{x}_2 ..... \mathbf{x}_N]$ (where N is the ensemble size) to estimate the mean and covariance. The analysis step of the EnKF at any particular time is given as

$$\mathbf{x}_i^{\mathrm{a}} = \mathbf{x}_i^{\mathrm{f}} + \mathbf{K}(\mathbf{y}_i - \mathbf{H}\mathbf{x}_i^{\mathrm{f}}), \qquad i = 1, 2, ..., N, \tag{1}$$

$$\mathbf{K} = \mathbf{P}^{\mathrm{f}}\mathbf{H}^{\mathrm{T}}(\mathbf{H}\mathbf{P}^{\mathrm{f}}\mathbf{H}^{\mathrm{T}} + \mathbf{R})^{-1}, \tag{2}$$

where $\mathbf{K}$ is referred to as the Kalman gain; $\mathbf{x}_i^{\mathrm{a}}$ and $\mathbf{x}_i^{\mathrm{f}}$ denote the ith analysis and forecast state member, respectively; $\mathbf{y}_i$ is the ith vector of perturbed observations; $\mathbf{H}$ is the observation operator, that is, mapping the state variable to the observation space (assumed linear here for simplification[1]); $\mathbf{P}^{\mathrm{f}}$ is the ensemble forecast-error covariance matrix and $\mathbf{R}$ is the observation-error covariance matrix. The superscripts "a", "f" and "T" stand for analysis, forecast and matrix transpose, respectively. For clarity, the time index is omitted from the notation. The term $(\mathbf{y}_i - \mathbf{H}\mathbf{x}_i^{\mathrm{f}})$ in Equation 1 is the discrepancy between the observations and the ensemble members, often referred to as the sample innovation. The ensemble forecast-error covariance matrix $\mathbf{P}^{\mathrm{f}}$ is never computed explicitly; however, it is decomposed as follows:

$$\mathbf{P}^{\mathrm{f}} = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{x}_i^{\mathrm{f}} - \bar{\mathbf{x}}^{\mathrm{f}})(\mathbf{x}_i^{\mathrm{f}} - \bar{\mathbf{x}}^{\mathrm{f}})^{\mathrm{T}} = \frac{1}{N-1}\mathbf{A}^{\mathrm{f}}(\mathbf{A}^{\mathrm{f}})^{\mathrm{T}}, \tag{3}$$

where $\bar{\mathbf{x}}^{\mathrm{f}}$ is the mean of the forecast ensemble:

$$\bar{\mathbf{x}}^{\mathrm{f}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i^{\mathrm{f}},$$

and $\mathbf{A}^{\mathrm{f}}$ is the ensemble anomalies matrix, the columns of which are the perturbations; that is, for $i = 1, ..., N$,

$$[\mathbf{A}^{\mathrm{f}}]_i = \mathbf{x}_i^{\mathrm{f}} - \bar{\mathbf{x}}^{\mathrm{f}}.$$

Similarly, the analysis-error covariance matrix can be computed from the ensemble of analysis states, but is not required in the implementation.

For hard data with known observational likelihood, each ensemble member is updated independently using observations that are perturbed with $\mathcal{N}(0, \mathbf{R})$ as shown in Burgers *et al.* (1998) and Evensen (2003). For observations with detection limits, the likelihood is truncated and is therefore non-Gaussian. How can this information be incorporated in an EnKF system?

[1]The proposed algorithm can be applied for nonlinear operators.

2.1.1 | Observations with a detection limit

When the observations have a detection limit, one may not have a full access to the observation likelihood. For simplicity, we will consider only the case with an upper detection limit on the observations, rather than a lower limit, without loss of generality. Observations with detection limit can be characterized into two parts:

1. hard data or in-range observations ($\mathbf{y}_{\text{ir}}$);
2. soft data or OR observations ($\mathbf{y}_{\text{or}}$), that is, no specific value of the observed quantity.

Within the Bayesian framework, the goal of DA is to estimate the posterior distribution of the model state. According to Bayes' rule, the posterior distribution $p(\mathbf{x}|\mathbf{y})$ is proportional to the product of a prior $p(\mathbf{x})$ and the observation likelihood $p(\mathbf{y}|\mathbf{x})$ as follows:

$$p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{x})p(\mathbf{x}). \tag{4}$$

For an observation with detection limit, Equation 4 can be split into two, depending on the nature of the observation, that is,

$$p(\mathbf{x}|\mathbf{y}) \propto \begin{cases} p(\mathbf{y}_{\text{ir}}|\mathbf{x})p(\mathbf{x}), & \text{when } \mathbf{y} = \mathbf{y}_{\text{ir}}, \text{ in-range observations,} \\ p(\mathbf{y}_{\text{or}}|\mathbf{x})p(\mathbf{x}), & \text{when } \mathbf{y} = \mathbf{y}_{\text{or}}, \text{ OR observations.} \end{cases} \tag{5}$$

Although we do not have an *a priori* OR observation likelihood $p(\mathbf{y}_{\text{or}}|\mathbf{x})$ to solve for the posterior $p(\mathbf{x}|\mathbf{y})$ given in Equation 5, one can always postulate an OR observation likelihood based on climatology or expert opinions. A detailed discussion about the choice of OR observation likelihood is given in section 2.2. In the following section, we will introduce the EnKF-SQ and present its implementation, along with its main differences from the PDEnKF.

2.2 | The ensemble Kalman filter semi-qualitative

The KF minimizes the forecast error variance and this is achieved by updating state variables, eventually moving them, on average, towards the observations. The update of the prior given in Equation 5 for in-range observations is straightforward. However, for OR observations it is not so clear, since we do not have the distribution of the observation. As such, an assumption about the OR observation likelihood, which should be physically consistent with the observed quantity and the qualitative information we have about it, is required to solve Equation 5.

2.2.1 | The partial deterministic EnKF

As a way to do that, Borup *et al.* (2015) proposed a DA scheme, namely PDEnKF, to solve the Bayesian system in Equation 5. The authors assumed the OR observation likelihood to be constant outside the observable range. Furthermore, the likelihood function inside the observable range is set to be determined by the in-range observation uncertainty, because measurement errors make it possible for in-range values to be wrongly observed as out of range. Inspired by the PDEnKF, we present an EnKF-SQ that uses a stochastic EnKF.

In contrast to the stochastic update, the PDEnKF follows Sakov and Oke (2008) and uses two different equations for updating the ensemble mean and anomalies, the anomalies being updated by half the gain in the form of implicit inflation. In some cases of partial update, the half-gain does not maintain the anomalies centered on the analysis mean: if the mean is within the range and the observation outside, the "half-gain" will leave the anomalies further inside the range than if the partial update were applied to the mean. Borup *et al.* (2015) have opted for this non-centered partial analysis scheme in order to maintain more ensemble spread.

In the EnKF-SQ, instead of a constant uniform OR observation likelihood, we propose to use a two-piece Gaussian distribution (Fechner, 1897; Gibbons and Mylroie, 1973) as the OR observation likelihood. In other words, the uniform likelihood of Borup *et al.* (2015) is replaced by a Gaussian distribution with varying observation-error variance outside the observable range.

A two-piece Gaussian distribution is obtained by merging two opposite halves of the two Gaussian probability densities (pdfs) at their common mode, given as follows:

$$f(x) = \begin{cases} W\exp\left[-\frac{(x-\mu)^2}{2\sigma_1^2}\right], & x \leq \mu, \\ W\exp\left[-\frac{(x-\mu)^2}{2\sigma_2^2}\right], & x > \mu, \end{cases} \tag{6}$$

where $W = \sqrt{\frac{2}{\pi}}(\sigma_1 + \sigma_2)^{-1}$ is a normalizing constant, μ is the common mean, and σ_1 and σ_2 are the standard deviations (std) of the two Gaussian pdfs. The common mean μ is located at the detection limit, as it is the last possible value the gauge could detect with known observation uncertainty. In essence, the common mean μ is nothing but the mode of a two-piece Gaussian distribution. Note that, for the function $f(x)$, the mean does not coincide with the mode, given the skewness of the distribution. The reasons for choosing a Gaussian likelihood, over a uniform one, in the unobservable range are as follows.

- OR observations do not give a specific value of the observed quantity, but an educated guess can always be made about a realistic range of values: for instance, using a climatology of the values in the unobservable range. Imposing a uniform density outside the observable range gives equal weight to all values until infinity, whereas extremely high values are usually less realistic in most applications, like wind speed and ice thickness among others.
- In order to implement the stochastic EnKF, one needs to perturb the observations (Equation 1) with a Gaussian distribution of covariance matrix $\mathbf{R}$. For the OR uniform likelihood this is technically impossible, since the uniform tail is not integrable. Even if the OR uniform likelihood

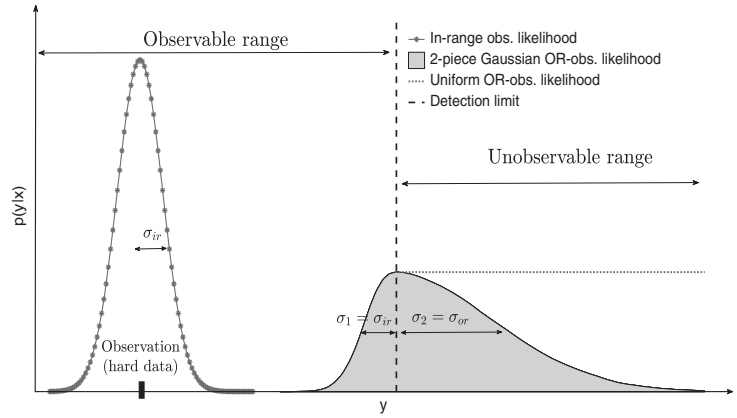


FIGURE 1 Illustration of the two-piece Gaussian OR observation likelihood, when a gauge has an upper observation limit. The in-range observation is shown by a small black rectangle and the corresponding Gaussian likelihood by a solid gray line. The two-piece Gaussian likelihood is plotted in gray and Uniform OR-obs likelihood from Borup *et al.* (2015) in dashed gray. σ_{ir} is the observation-error std for hard data and σ_{or} is the educated guess of the OR-obs error std

were limited to a finite upper bound, the choice of that upper bound would have to be justified by the nature of the variable. In principle, there is no restriction on the choice of OR likelihood probability distribution, but the Gaussian distribution has practically convenient properties for our purpose (the simulation does not generate excessive outliers and the Bayesian interpretation is relatively simple, see below).

In addition, we assume that the observation-error variance of the Gaussian half that is inside the observable range from the detection limit is equal to the in-range observation-error standard deviation (σ_{ir}), as in Borup *et al.* (2015). An example of the two-piece Gaussian OR observation likelihood having an upper detection limit is shown in Figure 1. As shown, the two-piece Gaussian likelihood is right-skewed, because of the higher OR observation-error standard deviation σ_{or}. Choosing a *proper* σ_{or} is very important, as it will be used to generate perturbations and thereby to update ensemble members. The choice should be consistent with the possible values in the unobservable range of the underlying observed variable.

2.2.2 | Choice of σ_{or}

The observation-error standard deviation for the Gaussian half outside the observable range (σ_{or}) is an arbitrary choice with different possibilities. If the pdf of the climatological data for the observed quantity is available, then σ_{or} can be approximated by using mean of out-of-range climatological values:

$$\sigma_{or} = -\mu + \left(\int_{\mu}^{+\infty} y f_{clim}(y)\, dy \right), \tag{7}$$

where $f_{clim}(y)$ is the pdf of the climatological data of the observed quantity, and μ is the detection limit point. The second term on the right-hand side of Equation 7 is the expectation of the climatological distribution for the values above the detection limit. Equation 7 is used to generate σ_{or} values in all of the experiments presented in section 3. Sensitivity experiments using different values for σ_{or} are also conducted (section 3.3.2). In the absence of climatology for the observed data, an educated guess can be used based on expert knowledge about σ_{or}, considering that extremely high values are less likely and vice versa for a lower detection limit.

2.2.3 | Bayesian representation

According to Bayes' rule, the posterior distribution is proportional to the product of the prior and observation likelihood functions (Equation 4). For hard data, the posterior is simply the product of two Gaussian distributions and it is Gaussian. For OR observations, it is the product of a Gaussian prior distribution and a two-piece Gaussian likelihood. This is nothing but the product of two Gaussian distributions (*the prior and each half of a two-piece Gaussian*) on either side of the detection limit and hence the posterior is a piecewise Gaussian distribution meeting at μ. The two Gaussians used in the likelihood have their mode at the detection limit; their multiplication by the Gaussian prior (which can have its mode either to the left (Figure 2a) or to the right (Figure 2b) side of the detection limit) will result in a shift of the modes of the two Gaussian pieces to the same side of the detection limit. The posterior will join the two Gaussian pieces at the detection limit, keeping one piece with a mode to one side and only a tail to the other side of the second piece. Hence it is a unimodal piecewise Gaussian, although not a two-piece Gaussian.

Figure 2 illustrates the update when the mode of the prior distribution is (a) inside and (b) outside the observable range. The curves for the two-piece Gaussian likelihood and

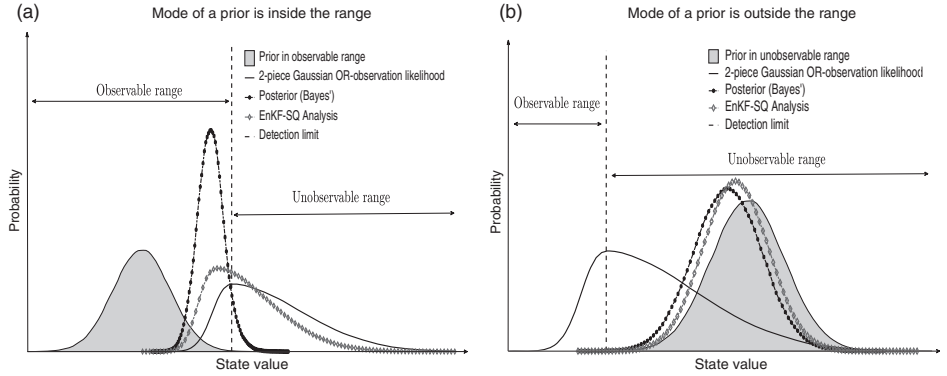


FIGURE 2 Bayesian posterior and EnKF-SQ analysis for a scalar update of a Gaussian prior with a two-piece Gaussian OR observation likelihood. (a) The mode of the prior is inside the observable range. (b) The mode of the prior is in the unobservable range

posterior Bayes distribution in Figure 2 are obtained by sampling the respective pdf with the inverse transform method. For each two-piece Gaussian-distributed random variate, we first generate a random number u from a uniform distribution $U[0, 1]$. Then the two-piece Gaussian-distributed random number is given as $x = F_X^{-1}(u)$, where F_X^{-1} is the inverse or quantile function of a two-piece Gaussian cumulative distribution function. The quantile function for a two-piece Gaussian cumulative distribution function is given as follows:

$$F_X^{-1}(u) = \begin{cases} \mu + \sigma_{\text{ir}}\Phi^{-1}\left(\frac{u}{W\sqrt{2\pi}\sigma_{\text{ir}}}\right), & \text{for } u \leq p = P[x \leq \mu], \\ \mu + \sigma_{\text{or}}\Phi^{-1}\left(\frac{u}{W\sqrt{2\pi}\sigma_{\text{or}}}\right), & \text{for } u > p = P[x \leq \mu], \end{cases}$$

and

$$P[x \leq \mu] = W\sqrt{2\pi}\sigma_{\text{ir}}\Phi\left(\frac{x-\mu}{\sigma_{\text{ir}}}\right),$$

where Φ and Φ^{-1} are a standard normal cumulative distribution function and its inverse, respectively. Φ^{-1} is obtained by a rational Chebyshev approximation. We have generated 100,000 samples to plot the figure.

As shown, when the mode of a prior distribution is inside the observable range (Figure 2a), the location of the posterior mode will be between the mode of the prior and the observation detection limit. This demonstrates the desired effect on the prior distribution by moving it towards the unobservable range. If the mode of the prior distribution is outside the observable range, then the update has a very small effect on it (Figure 2b), as expected, because of the high observation-error variance outside the observable range.

2.2.4 | Implementation and algorithm

The ensemble members can be seen as discrete samples of a continuous distribution. Updating the ensemble members given the hard data is performed by the stochastic EnKF, as in Equation 1. As for the soft data, Bayes' equation shows that ensemble members inside the observable range need to be updated towards the unobservable range. Intuitively, the ensemble members that lie in the unobservable range should be left untouched, as we do not have a specific value of the observation.

The proposed approach to update the ensemble for soft data is divided into two cases, depending whether the observed ensemble members, that is, $\mathbf{H}\mathbf{x}_i^{\text{f}}$, are inside or outside the observable range. Members inside (respectively outside) the observable range should be updated linearly with observation uncertainty σ_{ir} (respectively σ_{or}). During the update, the observation perturbations $\mathbf{y}_i$ can be generated by the inverse-transform method as described in section 2.2.3.

The Kalman gain $\mathbf{K}$ for a forecast ensemble member inside the observable range is calculated with the in-range observation-error standard deviation σ_{ir}. For a scalar case, this becomes $K_{\text{ir}} = \sigma_{\text{b}}^2\left(\sigma_{\text{b}}^2 + {\sigma_{\text{ir}}}^2\right)^{-1}$. If the member is outside the observable range, then the Kalman gain $\mathbf{K}$ is calculated with the out-of-range observation error standard deviation σ_{or}, that is, $K_{\text{or}} = \sigma_{\text{b}}^2\left(\sigma_{\text{b}}^2 + \sigma_{\text{or}}^2\right)^{-1}$. Here, the forecast error variance is denoted by σ_{b}^2. The rationale for using different observation-error variances, rather than a single one for the two-piece Gaussian distribution, is to be consistent with the Bayesian update represented in Figure 2: when a prior is inside the observable range, the posterior is the product of the prior and the portion of the two-piece Gaussian likelihood located below the detection limit. The same goes for the posterior part in the unobservable range. In other words, a prior inside (outside) the observable range is updated using the observation-error standard deviation σ_{ir} (σ_{or}). If, however, the variance of the two-piece Gaussian distribution had been used instead, a value between σ_{ir}^2 and σ_{or}^2, the algorithm would then update the in-range members too weakly and the OR members too strongly.

In the multivariate case, this can be achieved by simply changing the values of the observation-error variance to ${\sigma_{\text{ir}}}^2$

or σ_{or}^2 for an OR observation in the error covariance matrix $\mathbf{R}$, depending on the location of the forecast ensemble member. Note that the proposed algorithm for EnKF-SQ only supports uncorrelated observations i.e. matrix $\mathbf{R}$ is diagonal. If the observation errors are correlated, one can decorrelate them (Evensen, 2004) and proceed with the algorithm. An algorithmic implementation of the EnKF-SQ analysis is presented below:

Algorithmic steps

For an efficient processing of the update Equation 5, observations are preprocessed serially to sort out hard data (y_{ir}) and soft data (y_{or}) before proceeding to the analysis. The subscripts "ir" and "or" stand for the index number of any hard and soft data in observation vector $\mathbf{y}$, respectively.
For each ensemble member i:

1. For each OR observation y_{or}, apply the "or"th observation operator row $\mathbf{H}_{or}$ to ensemble member $\mathbf{x}_i^f$, to check whether the member is outside or inside the observable range.
2. Perform the operation below for all OR observations, in order to set the values of the observation-error variance in matrix $\mathbf{R}$ depending on the location of $\mathbf{H}_{or}\mathbf{x}_i^f$.
   Pseudo-code:
   for each OR observation y_{or}
   if $\mathbf{H}_{or}\mathbf{x}_i^f > \mu$
   $\mathbf{R}_{(or,or)} = \sigma_{or}^2$
   else
   $\mathbf{R}_{(or,or)} = {\sigma_{ir}}^2$
   end if
   end for each $\mathbf{y}_{or}$
3. Calculate the Kalman gain matrix $\mathbf{K}$ with the updated $\mathbf{R}$.
4. Update the forecast ensemble member $\mathbf{x}_i^f$ using EnKF update Equation 1, where the perturbation vector $y_{i,or}$ and $y_{i,ir}$ are generated from the two-piece Gaussian likelihood and $\mathcal{N}(y_{ir}, {\sigma_{ir}}^2)$ for OR and in-range observations respectively.
5. Repeat the process for all N ensemble member $\mathbf{x}_i^f$ to obtain the analysis ensemble.

End the loop on i.

A flowchart for the EnKF-SQ update scheme is given in Figure 3.

To study the posterior obtained by the proposed EnKF-SQ algorithm, we superimpose the EnKF-SQ analysis on the Bayesian solution in both panels of Figure 2. The EnKF-SQ analysis is obtained from the exact same prior and likelihood as in section 2.2.3. Since the likelihood is not Gaussian, we do not expect the EnKF-SQ ensemble to coincide with the Bayesian solution. We used 10,000 ensemble members to sample the prior and two-piece Gaussian OR observation likelihood.

When the mode of a prior is outside the observable range, the EnKF-SQ scheme yields approximately the same posterior as that of Bayes' rule (Figure 2b), but marginally closer to the prior distribution. The EnKF-SQ slightly *underassimilates* in this case, which conforms with the intention of little impact of OR observations on OR forecast members.

In contrast, when the mode of the prior is inside the observable range, the posterior obtained by Bayes' rule has a sharper peak, whereas the analysis obtained from the EnKF-SQ has a thick tail in the OR domain (Figure 2a). The large deviation from the Bayesian solution is a sign of the sensitivity of the linear EnKF-SQ update to a skew input likelihood. This skewness was already present in the previous case; however, it was not visible with larger OR observation errors. In the present case, the posterior EnKF-SQ ensemble is closer to the likelihood than the Bayesian solution, so it can be stated that the EnKF-SQ *overassimilates* in this case, although it does return a larger ensemble spread than the Bayesian solution, which may be counterintuitive for EnKF practitioners. It is worth noticing that the posterior modes obtained from Bayes' rule and EnKF-SQ analysis still remain close to each other, as intended in Borup *et al.* (2015).

Note that the posterior represented in Figure 2 may not be very well sampled in practice if the ensemble is very small. The inconvenience of sampling errors and skewness will be evaluated with toy models in the following sections.

3 | NUMERICAL TESTS

In this section we present and analyze the assimilation results obtained using the proposed EnKF-SQ algorithm. We use two different toy models to test and evaluate the behavior of the EnKF-SQ. The first is a linear subsurface flow model (LSST) and the second is the nonlinear Lorenz-40 (L40) model of Lorenz and Emanuel (1998). We conduct various sensitivity experiments with variable ensemble size, detection limit, and σ_{or}. We also compare the performance of the EnKF-SQ against the PDEnKF and with two different versions of the stochastic EnKF, denoted as follows.

1. EnKF-ALL: No observation detection limit is applied during DA experiments, thus all observations are hard data.
2. EnKF-IG: Assimilating only hard data and ignoring soft data during the analysis. The goal for testing with EnKF-IG is to assess the added information introduced by EnKF-SQ.

First we give a brief description of the models and the configuration used in the tests, and then we discuss the results from different numerical experiments.

3.1 | The linear subsurface transport (LSST) model

We consider a 1D subsurface transport model in an unconfined aquifer. The transport model is driven by a steady subsurface flow using a combined Darcy's law and continuity

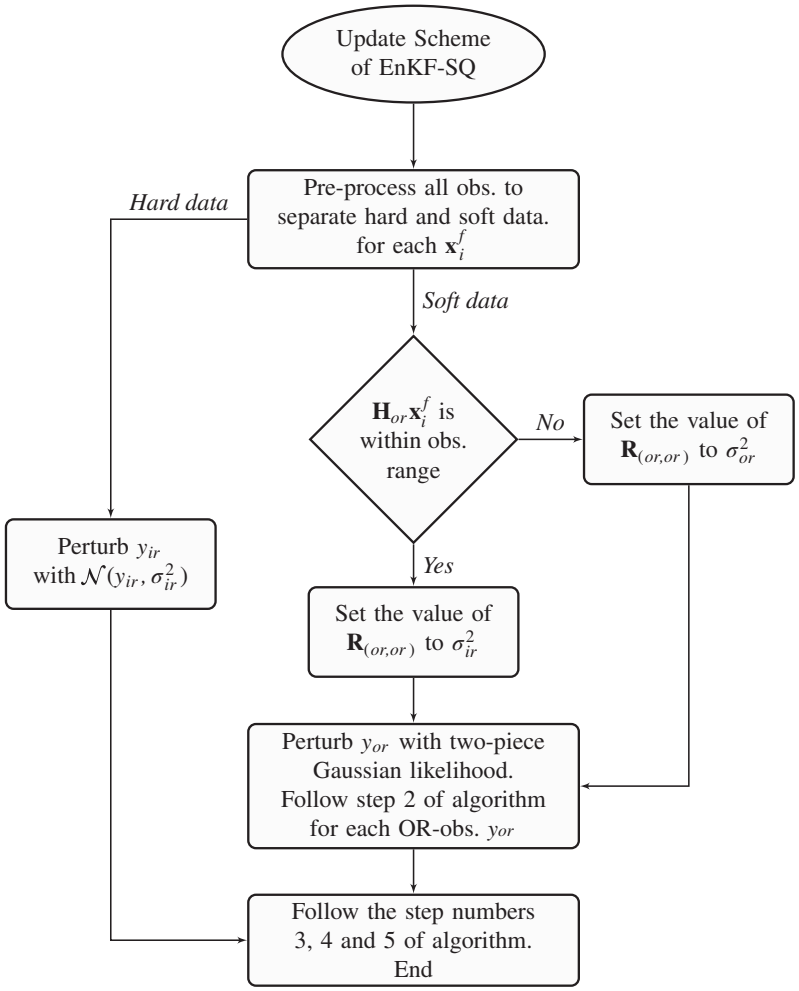


FIGURE 3 Flowchart for implementation of the EnKF-SQ. Note that the algorithm does not lend itself to matrix multiplications, as in Evensen (2003)

equation. Groundwater flows from west to east at a reference Darcy velocity of 1.18×10^{-4} m/s. Periodic water-head boundary conditions are assumed. The domain is uniformly discretized into 100 cells, with each cell measuring 10 m in length.

The generalized 1D linear solute transport model is obtained from mass conservation of species, defined as

$$r_c \frac{\partial(\phi C)}{\partial t} + \frac{\partial(UC)}{\partial x} = q, \tag{8}$$

where r_c is the retardation coefficient, ϕ is the porosity, t is time (s), C is the concentration of contaminant species (ppm), U is the Darcy velocity (m/s), and q is the contaminant source. An initial condition for the concentration is specified: $C(x, 0) = 3+\sin(5x_i)$, where x_i is the length of the ith grid cell. The time step is set to 10 hr. The porosity is uniform and equal to 33.4%, with a retardation coefficient of 5.19. The contaminant source, q, at every time step is equal to 3×10^{-6} ppm. Water flowing from the western boundary is contaminated with 5 ppm concentration value. Using these parameters, a reference run solution is simulated for a period of 4 years. In order to mimic realistic scenarios, we impose model error in the forecast model. Essentially, we perturb the transport parameters such that $\phi = 30\%$ and $r_c = 6.87$. We also add Gaussian noise $\mathcal{N}(0, 0.01)$ to the contaminant source and the Darcy velocity field.

3.2 | The L40 model

The L40 model (Lorenz and Emanuel, 1998) is a chaotic and nonlinear model with 40 state variables. It imitates the evolution of an unspecified scalar meteorological quantity, for instance temperature or vorticity along a latitude circle. This model has been used for testing ensemble-based assimilation methods in a number of earlier studies (Anderson, 2001; Whitaker and Hamill, 2002; Sakov and Oke, 2008). The model assumes cyclic boundary conditions as follows:

$$\frac{dz_i}{dt} = (z_{i+1} - z_{i-2})\, z_{i-1} - z_i + F, \quad i = 1, ..., 40; \quad (9)$$

$$z_0 = z_{40}, \quad z_{-1} = z_{39}, \quad z_{41} = z_1,$$

where z_i is the ith state variable and F is a forcing term. The time step is set to $\Delta t = 0.05$ units (that is, 6 hr in real atmospheric time). The model is integrated forward in time using fourth-order Runge–Kutta. The reference (truth) trajectory is initialized by setting $F = 8$, $z_i = F$, $\forall i \neq 20$, and $z_{20} = F + 0.001$. A reference run solution is simulated for a period of 5 years (7300 steps). Initial ensemble members are obtained by perturbing the mean state of the reference trajectory with Gaussian noise, $\mathcal{N}(0, 3)$. Observations are collected from the reference trajectory and then contaminated using a Gaussian distribution $\mathcal{N}(0, 1)$. We impose a model error for data assimilation experiments by changing the forcing parameter to $F = 8.1$.

3.3 | Results

Experiments are performed over periods of 5 and 4 years for the L40 and LSST models, respectively. The size of the ensemble is chosen based on a series of sensitivity experiments, and is set to 75 and 30 for the L40 and LSST models, respectively. The choice is made so that tuning parameters such as inflation and localization are not needed. The goal is to assess the performance of the EnKF-SQ, PDEnKF, and EnKF-IG schemes for a large enough ensemble, without the necessity to mitigate sampling errors and other filter-related deficiencies. For the L40 model, all 40 variables are observed and assimilated every day (that is, every fourth time step). In the LSST model, 80 variables are observed, with a regularly spaced observing network, every tenth time step.

The forecast root-mean-square error (RMSE) is used to evaluate the filter performance. Given the n-dimensional mean forecast state vector $\hat{\mathbf{x}}_t^{\mathrm{f}} = (\hat{x}_{t,1}^{\mathrm{f}}, \hat{x}_{t,2}^{\mathrm{f}}, ..., \hat{x}_{t,n}^{\mathrm{f}})$ at time t, and if $t_{\max}$ is the final time, then the time-averaged RMSE is defined as

$$\widehat{RMSE} = \frac{1}{t_{\max}} \sum_{t=1}^{t_{\max}} \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{x}_{t,i}^{\mathrm{f}} - x_{t,i}^{\mathrm{r}})^2}, \quad (10)$$

where $\mathbf{x}_t^{\mathrm{r}} = (\hat{x}_{t,1}^{\mathrm{r}}, \hat{x}_{t,2}^{\mathrm{r}}, ..., \hat{x}_{t,n}^{\mathrm{r}})$ is the reference state vector at time t. Each filter run is then repeated $L = 10$ times, with different random seeds to initialize the random number generator. The average RMSE over these L runs is then reported as

$$\overline{RMSE} = \frac{1}{L} \sum_{l=1}^{L} \widehat{RMSE}_l. \quad (11)$$

3.3.1 | General behavior of the EnKF-SQ

Figure 4 shows the time evolution of the RMSE and average ensemble spread (AES) of forecast ensemble members obtained using the EnKF-SQ, PDEnKF, and EnKF-IG. Generally, for a "healthy" assimilation framework, the RMSE is expected to match the AES plus the observation errors. We set different detection limits on observation in both models, such that on average 80% of observations fall out of range, that is, they become soft data. We also show the RMSE of a free run (no DA) in both models. For clarity, we superimpose the moving average of RMSE and AES of all three schemes in both panels of Figure 4. As shown in Figure 4, assimilating soft data using the EnKF-SQ improves the forecast RMSE in both models. As shown, among the three tested filters the EnKF-IG is the least accurate. Clearly, assimilating fewer data degrades the quality of the forecast. We note that the RMSE and total spread (AES + observation-error standard deviation) are of the same order, indicating no signs of inbreeding or divergence. As shown, both the EnKF-SQ and the PDEnKF benefit from assimilating soft data. On average, the proposed EnKF-SQ estimates are 20 and 12% more accurate than those of the PDEnKF for the LSST and L40 models, respectively (Figure 4).

To visualize the time evolution of the ensemble for the EnKF-SQ, PDEnKF, and EnKF-IG with the LSST model, we plot the concentration of a randomly chosen observed and unobserved state variable versus time in Figure 5. Analyzing the results from Figure 5 along with Figure 4 demonstrates clearly that assimilating soft data not only improves the RMSE but also reduces the uncertainty in the forecast, by shrinking the ensemble spread around the truth. As expected, the EnKF-IG is the least accurate, generating low concentration values when the observations are above the threshold, which the EnKF-SQ avoids successfully. Compared with the PDEnKF, the proposed scheme matches the truth trajectory better. This can be clearly observed for the time intervals (3000, 3700). Similar behavior was also observed for the L40 estimates (not shown).

3.3.2 | Sensitivity experiments

Sensitivity experiments are conducted by varying both the ensemble size and detection limits using the L40 model and the results are presented in Figure 6. The RMSE values obtained for these experiments are averaged overa 5 year-long DA run. The goal is to assess the convergence rate of the EnKF-SQ, while increasing N from 25 to 150. The resulting RMSE is plotted against $1/\sqrt{N}$, given that the precision of Monte Carlo methods varies as a function of $1/\sqrt{N}$.

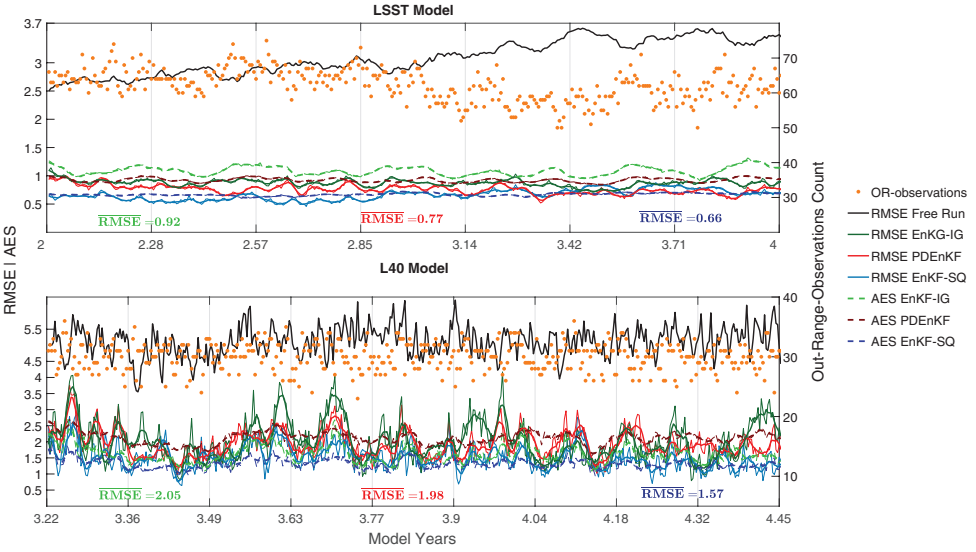


FIGURE 4 Time evolution of the of RMSE (solid lines), AES (dashed lines) and their moving averages in thicker solid lines. The orange dots represent the number of OR observations during the assimilation time. The top and bottom panels show EnKF-SQ, PDEnKF, and EnKF-IG results from the LSST and L40 models, respectively [Colour figure can be viewed at wileyonlinelibrary.com]

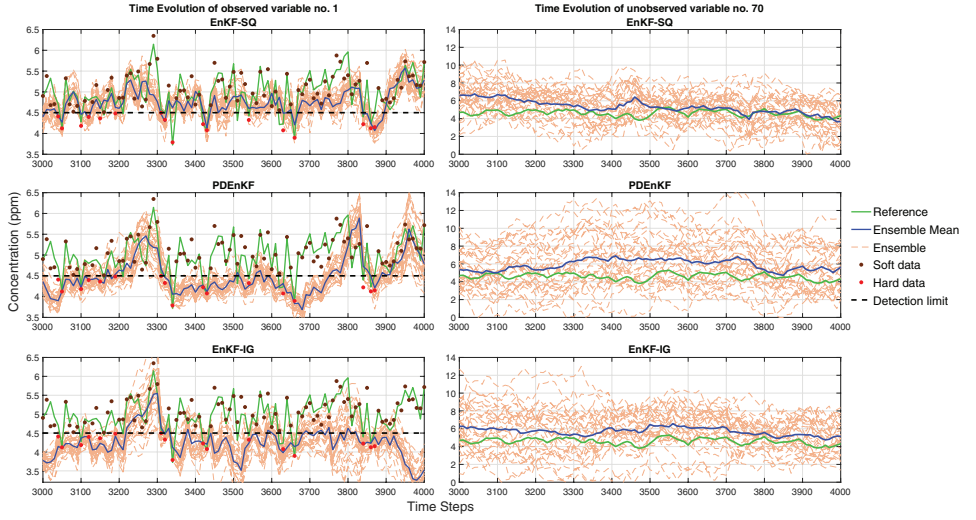


FIGURE 5 Time evolution of the forecast ensemble (dashed orange), ensemble mean (solid blue), and truth (solid green) for observed and unobserved state variable numbers 1 (left panels) and 70 (right panels), respectively in the LSST model, obtained using EnKF-SQ (top panels), PDEnKF (middle), and EnKF-IG (bottom) [Colour figure can be viewed at wileyonlinelibrary.com]

As the ensemble size increases, the RMSE value for each scheme naturally decreases, as shown in Figure 6a, although none of them is linear in $1/\sqrt{N}$. For all tested ensemble sizes, the proposed scheme is consistently more accurate than the EnKF-IG. The PDEnKF appears to benefit from small ensemble sizes (e.g. 25 and 35) and outperforms the

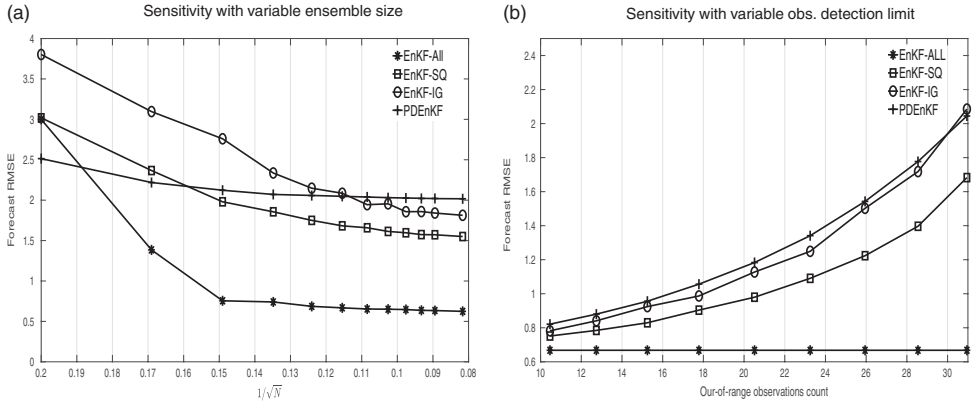


FIGURE 6 Root-mean-square error of the forecast estimates resulting from the EnKF-All, EnKF-SQ, EnKF-IG, and PDEnKF for sensitivity experiments with varying ensemble size ($N = 25, 35, 45, \ldots, 150$) in the left panel and the observation detection limit ($N = 75$) in right panel, for the L40 model. For the sensitivity experiment with ensemble size (left), approximately 80% (≈ 32) of total observations are OR observations

EnKF-SQ. It is worth noticing that for small ensemble sizes the EnKF-SQ performance is as good as the EnKF-ALL, and for $N = 25$ the estimates of both schemes overlap. We also note that PDEnKF performs better than EnKF-ALL for an ensemble size of 25, which could be related to the fact that it is a deterministic filter. Various studies (e.g. Sakov and Oke, 2008) have reported that square-root deterministic methods handle sampling errors better than the stochastic EnKF for small ensemble size.

Changing the detection limit on observations is done such that the number of observations falling out of range increases gradually and the system has fewer hard data to assimilate. The forecast RMSE resulting from the EnKF-SQ is shown in Figure 6b to vary between two extreme cases, that is, EnKF-IG and EnKF-ALL. The forecast RMSE increases more slowly with more numerous OR observations with the EnKF-SQ than the PDEnKF, indicating more robust performance in difficult cases with few hard data. Even with very few hard data to assimilate, the EnKF-SQ estimates are almost 19% more accurate than those of the EnKF-IG. The PDEnKF's performance becomes similar to the EnKF-IG, having no benefit of assimilating qualitative information. All four schemes converge towards approximately the same RMSE as more hard data are assimilated.

By comparison with the EnKF, σ_{or} is the only new parameter introduced in the EnKF-SQ. This imposes only minor changes to existing EnKF codes. We perform sensitivity experiments by introducing a scalar multiplier to Equation 7, namely α, to examine the behavior of the EnKF-SQ and the impact of using more skewed ensembles. We vary α between 0.05 and 1.85 with a step size of 0.15. The new form of Equation 7 is shown below:

$$\sigma_{\text{or}^*} = \underbrace{\alpha\left[-\mu + \left(\int_{\mu}^{+\infty} y f_{\text{clim}}(y)dy\right)\right]}_{\sigma_{\text{or}}}. \tag{12}$$

The reason for choosing 1.85 as the upper bound for α values is the natural range of variability of the Lorenz system. Concerning the L40 experiments, when α increases beyond 1.85, the value of OR observation-error standard deviation (σ_{or}) becomes large enough that it generates perturbed observations exceeding the natural variability of the model. Hence, assimilating such observations with the in-range observation-error standard deviation (σ_{ir}) gives unrealistic values for the model state, which may be a limitation of the Gaussian likelihood. The LSST model has no such restriction, but the same α interval has been used for a clear comparison of the two models.

We plot the RMSE values of the analysis states, from both models, in addition to the absolute skewness of the analysis and observation likelihood versus α for the EnKF-SQ in Figure 7. The values obtained by the PDEnKF are independent of α and shown for reference. The skewness of analysis and observation likelihood are evaluated as the average absolute value of each variable skewness and only at the last assimilation step:

$$\text{skew}_{\text{a}} = \frac{1}{n}\sum_{j=1}^{n}\left|\frac{\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{x}^{\text{a}}_{t_{\max},j,i} - \hat{\mathbf{x}}^{\text{a}}_{t_{\max},j}\right)^3}{\left(\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{x}^{\text{a}}_{t_{\max},j,i} - \hat{\mathbf{x}}^{\text{a}}_{t_{\max},j}\right)^2\right)^{3/2}}\right|, \tag{13}$$

$$\text{skew}_{\text{o}} = \frac{1}{m}\sum_{j=1}^{m}\left|\frac{\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{y}_{t_{\max},j,i} - \hat{\mathbf{y}}_{t_{\max},j}\right)^3}{\left(\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{y}_{t_{\max},j,i} - \hat{\mathbf{y}}_{t_{\max},j}\right)^2\right)^{3/2}}\right|, \tag{14}$$

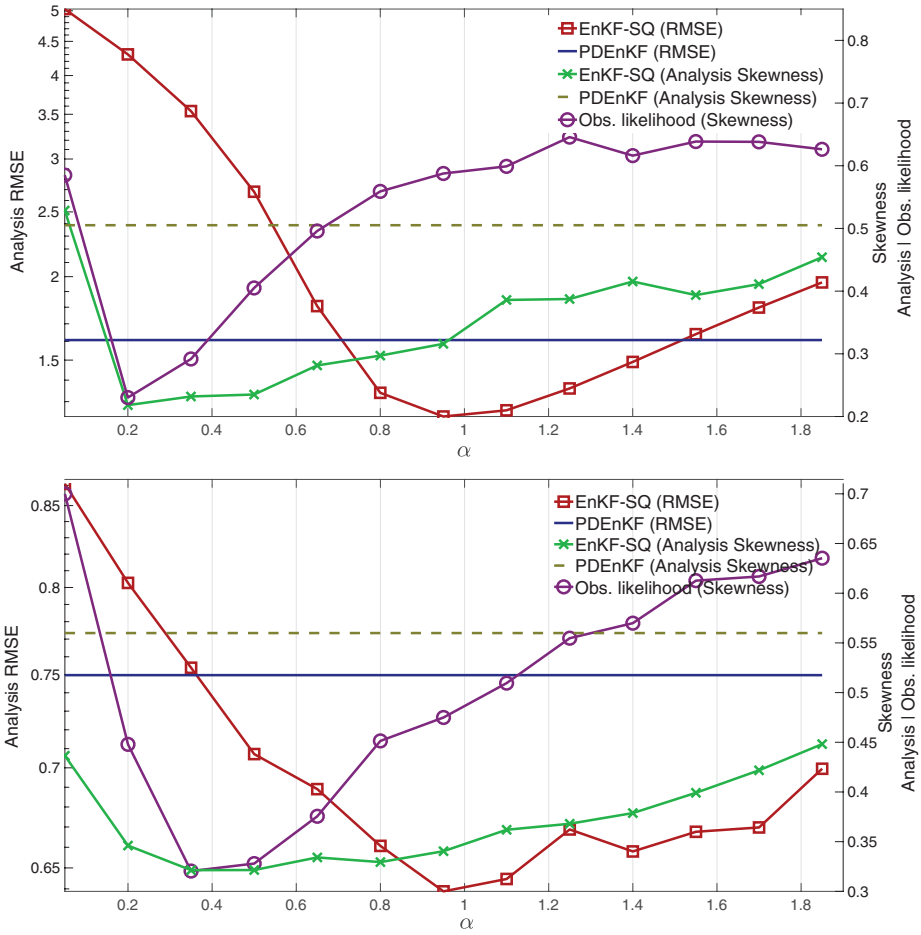


FIGURE 7 Performance of the EnKF-SQ and PDEnKF using 30 and 75 ensemble members, for LSST (top panel) and L40 (bottom panel) with increasing values of α. The analysis RMSE is demonstrated, in addition to the skewness of the posterior and likelihood distributions [Colour figure can be viewed at wileyonlinelibrary.com]

where m is the number of observations, $\mathbf{x}^{a}_{t_{\max},j,i}$ and $\mathbf{y}_{t_{\max},j,i}$ are the ith analysis ensemble member and ith observation perturbation vector at time $t_{\max}$, respectively, and $\hat{\mathbf{x}}^{a}_{t_{\max},j,i}$ and $\hat{\mathbf{y}}_{t_{\max},j,i}$ are the analysis mean and mean of the observation perturbation vector at time $t_{\max}$, respectively.

RMSE changes in Figure 7 indicate that when the value of α approaches 1, that is, close to nominal value of σ_{or} in Equation 7, the EnKF-SQ outperforms the PDEnKF. As the value of α moves away from 1, the performance of EnKF-SQ starts to deteriorate, especially for the L40 model. This, however, is less obvious for the LSST model. This can be explained by a poor sampling of the two-piece Gaussian likelihood using a finite ensemble size when σ_{or} is assigned very high and/or low values. For instance, in the case of a high σ_{or}, sampling might produce very large perturbations of observations (outliers), which can make the analysis increments more erratic. On the other hand, small values of σ_{or} are also detrimental, as they are prone to generate samples concentrated around the detection limit, thus pulling the analysis close to an artificial threshold limit. This confirms a posteriori the choice of the nominal value of σ_{or} and the importance of a good knowledge of climatological values: if the climatological average of L40 OR values is biased by more than 50% (α lower than 0.5 or larger than 1.5), then the flat likelihood of the PDEnKF makes a better option. The linear LSST model is more permissive in this respect, since the EnKF-SQ will beat the PDEnKF even with values of α more than 100% off the nominal value. This could be because non-Gaussianity

is reduced in a linear model like the LSST (central limit theorem), in contrast to the nonlinear and chaotic L40 model.

Figure 7 also shows the absolute skewness of the observation likelihood and analysis ensemble. The skewness of the EnKF-SQ analysis ensemble follows the same trend as that of the likelihood, though the linear EnKF-SQ update makes it less skewed. To illustrate, as α increases, the observation likelihood transitions from being right to left skewed (not shown). Likewise, the analysis follows a similar behavior. The analysis ensembles are quite severely skewed (typical skewness is from 0.3–0.5 for the EnKF-SQ and higher with the PDEnKF update scheme), which does not seem to affect the EnKF-SQ performance directly. The minimum RMSE does not even coincide with the minimum skewness. This indicates that the method can handle some degree of non-Gaussianity, which makes it useful for assimilating soft data with the EnKF-SQ and PDEnKF.

4 | SUMMARY AND DISCUSSION

In practice, many observations are only available within a confined range. Qualitative information measured above or below the detection limit can still be exploited by data assimilation, although current methods only consider hard data. In this article, we proposed a new DA algorithm, referred to as EnKF-SQ, in order to assimilate semi-qualitative observations through an explicit treatment of soft data. The update algorithm requires a preprocess step, in which observations are split into two groups, hard and soft data. This is then followed by an update of the forecast ensemble using the Kalman update. An assumption is imposed that the observation likelihood should be a two-piece Gaussian and the mode of the likelihood is positioned at the detection limit. Members falling inside or outside the observable range are then separated, to achieve consistent update by soft data. This makes it necessary to update each forecast ensemble member individually, but not in parallel. Computationally, this is not a major issue, as in many applications the update only represents a few per cent of the costs of the ensemble propagation step (Sakov *et al.*, 2012) and a local EnKF-SQ would still run local updates in a parallel loop.

The new EnKF-SQ has been evaluated in linear subsurface transport and nonlinear Lorenz-40 models. Its performance has been compared for two different versions of the stochastic EnKF, namely EnKF-ALL (no detection limit on observations) and EnKF-IG (no assimilation of soft data), in addition to the previously introduced partial deterministic ensemble Kalman filter (PDEnKF), which is built over a deterministic EnKF and uses a uniform OR prior likelihood. Our numerical results suggest that assimilating soft data with the EnKF-SQ improves the overall forecast accuracy. The scheme outperforms the EnKF-IG, with reasonable computing time and ensemble sizes lower than 100 for systems of dimension greater than 20. Thus it does not suffer from the curse of dimensionality. This suggests that EnKF-SQ is a viable method and can be implemented with more realistic applications of the EnKF.

Sensitivity experiments on the chosen value of OR observation likelihood error variance σ_{or} imply that, if chosen *properly*, the EnKF-SQ performs better than the PDEnKF. This may not be true for all types of application, however, because the differences in performance are small and some observations may be represented by an OR likelihood with a fatter tail than the Gaussian distribution. Such cases are not addressed in the present work. As far as the two-piece Gaussian likelihood goes, we found that, even though it might increase the skewness of the posterior distribution, the benefits of assimilating soft data outweigh the inconvenience of non-Gaussianity in both linear and nonlinear cases.

The question arises as to whether the assimilation of an arbitrary value in the out-of-range domain will perform as well as the EnKF-SQ, with less algorithmic complexity. This has not been tested, but we note that assimilating a hard pseudo-observation in the OR domain would not introduce asymmetric information as the EnKF-SQ does, so the approach would unnecessarily update forecast ensemble members that fall rightly in the OR domain.

Is this semi-qualitative approach applicable to other data assimilation methods? It requires a stochastic data assimilation method to treat the ensemble members as possible realizations of the underlying random variables. Extensions to deterministic methods are therefore not straightforward. The link to optimal interpolation (OI) can be made by geostatistical methods through randomization (Emery and Robles, 2008), but this would make the OI method much more costly. The extension from ensemble filters to ensemble smoothers should, however, be straightforward.

ACKNOWLEDGEMENTS

The Authors thank Morten Borup for interesting discussions and hosting AS at DTU. We also thank François Counillon for insightful scientific discussions and Alberto Carrassi for suggesting the name “EnKF-SQ”. We are grateful to the two anonymous reviewers and handling editor M. Bocquet for insightful comments that have helped improve the manuscript. The research is funded by the Nordic Center of Excellence EmblA (Ensemble-based data assimilation for environmental monitoring and prediction) under NordForsk contract number 56801.

REFERENCES

Anderson, J.L. (2001) An ensemble adjustment Kalman filter for data assimilation. *Monthly Weather Review*, 129, 2884–2903.

Bocquet, M., Pires, C.A. and Wu, L. (2010) Beyond Gaussian statistical modeling in geophysical data assimilation. *Monthly Weather Review*, 138(8), 2997–3023.

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

Burgers, G., Jan van Leeuwen, P. and Evensen, G. (1998) Analysis scheme in the ensemble Kalman filter. *Monthly Weather Review*, 126, 1719–1724.

Chiles, J. and Delfiner, P. (1999) *Geostatistics: Modeling Spatial Uncertainty*. Wiley, New York, 695 p. https://doi.org/10.1002/9780470316993.

Daley, R. (1993) *Atmospheric Data Analysis*, 1st edition. Cambridge, UK: Cambridge University Press.

Emery, X. and Robles, L.N. (2008) Simulation of mineral grades with hard and soft conditioning data: application to a porphyry copper deposit. *Computational Geosciences*, 13(1), 79. https://doi.org/10.1007/s10596-008-9106-x.

Evensen, G. (2003) The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean Dynamics*, 53, 343–367.

Evensen, G. (2004) Sampling strategies and square root analysis schemes for the EnKF. *Ocean Dynamics*, 54, 539–560.

Fechner, G. (1897). In: Lipps, G.F. (Ed.) *Auftrage der Königlich Sächsichen Gesellschaft der Wissenschaften*. (in German). Engelmann: Leipzig, pp. 55–83.

Gharamti, M.E., Hoteit, I. and Sun, S. (2012) Low-rank Kalman filtering for efficient state estimation of subsurface advective contaminant transport models. *Journal of Environmental Engineering*, 138, 446–457. https://doi.org/10.1061/(ASCE)EE.1943-7870.0000484.

Ghil, M. and Malanotte-Rizzoli, P. (1991) Data assimilation in meteorology and oceanography. *Advances in Geophysics*, 33, 141–266.

Gibbons, J. and Mylroie, S. (1973) Estimation of impurity profiles in ion implanted amorphous targets using joined half Gaussian distributions. *Applied Physics Letters*, 22, 568–569. https://doi.org/10.1063/1.1654511.

Hornung, R.W. and Reed, L.D. (1990) Estimation of average concentration in the presence of nondetectable values. *Applied Occupational and Environmental Hygiene*, 5, 46–51.

Kaleschke, L., Maaß, N., Haas, C., Hendricks, S., Heygster, G. and Tonbøe, 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. https://doi.org/10.5194/tc-4-583-2010.

Kaleschke, L., Tian-Kunze, X., Maaß, N., Mäkynen, M. and Drusch, M. (2012) Sea ice thickness retrieval from SMOS brightness temperatures during the Arctic freeze-up period. *Geophysical Research Letters*, 39(5), L05501. https://doi.org/10.1029/2012GL050916.

Kalman, R.E. (1960) A New Approach to Linear Filtering and Prediction Problems ASME. *Journal of Basic Engineering*, 82(1), 35–45. https://doi.org/10.1115/1.3662552.

Kalnay, E. (2003) *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge, UK: Cambridge University Press.

Li, X. and Cheng, G. (1999) A GIS-aided response model of high-altitude permafrost to global change. *Science in China Series D: Earth Sciences*, 42, 72–79.

Lorenz, E.N. and Emanuel, K.A. (1998) Optimal sites for supplementary weather observations: simulation with a small model. *Journal of the Atmospheric Sciences*, 55, 399–414.

Reul, N., Tenerelli, J., Chapron, B., Vandemark, D., Quilfen, Y. and Kerr, Y. (2012) SMOS satellite L-band radiometer: a new capability for ocean surface remote sensing in hurricanes. *Journal of Geophysical Research: Oceans*, 117, C02006. https://doi.org/10.1029/2011JC007474.

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

Sakov, P., Counillon, F., Bertino, L., Lisæter, K., Oke, P. and Korablev, A. (2012) TOPAZ4: an ocean-sea ice data assimilation system for the north Atlantic and Arctic. *Ocean Science*, 8, 633.

Talagrand, O. (1997) Assimilation of observations, an introduction (special issue data assimilation in meteology and oceanography: theory and practice). *Journal of the Meteorological Society of Japan Series II*, 75(1B), 191–209.

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

Tippett, M.K., Anderson, J.L., Bishop, C.H., Hamill, T.M. and Whitaker, J.S. (2003) Ensemble square root filters. *Monthly Weather Review*, 131, 1485–1490.

Whitaker, J.S. and Hamill, T.M. (2002) Ensemble data assimilation without perturbed observations. *Monthly Weather Review*, 130, 1913–1924.

How to cite this article: Shah A, El Gharamti M, Bertino L. Assimilation of semi-qualitative observations with a stochastic ensemble Kalman filter. *Q J R Meteorol Soc.* 2018;1–13. https://doi.org/10.1002/qj.3381

Paper II

Assimilation of semi-qualitative sea ice thickness data with the EnKF-SQ

Shah, A., Bertino, L., Counillon, F., El Gharamti, M., and Xie, J.
Tellus A: Dynamic Meteorology and Oceanography, submitted

Assimilation of semi-qualitative sea ice thickness data with the EnKF-SQ

Abhishek Shah[1,*], Laurent Bertino[1], François Counillon[1], Mohamad El Gharamti[2], and Jiping Xie[1]

[1]Nansen Environmental and Remote Sensing Center, Bergen, Norway
[2]National Center for Atmospheric Research, Boulder, Colorado

Abstract

A newly introduced stochastic data assimilation method, the Ensemble Kalman Filter Semi-Qualitative (EnKF-SQ) is applied to a realistic coupled ice-ocean model of the Arctic, the TOPAZ4 configuration, in a twin experiment framework. The method is shown to add value to range-limited thin ice thickness measurements, as obtained from passive microwave remote sensing, with respect to more trivial solutions like neglecting the out-of-range values or assimilating climatology instead.

Some known properties inherent to the EnKF-SQ are evaluated: the tendency to draw the solution closer to the thickness threshold, the skewness of the resulting analysis ensemble and the potential appearance of outliers. The experiments show that none of these properties prove deleterious in light of the other sub-optimal characters of the sea ice data assimilation system used here (non-linearities, non-Gaussian variables, lack of strong coupling). The EnKF-SQ has a single tuning parameter that is adjusted for best performance of the system at hand. The sensitivity tests reveal that the results do not depend critically on the choice of this tuning parameter. The EnKF-SQ makes overall a valid approach for assimilating semi-qualitative observations into high-dimensional nonlinear systems.

Keywords— Semi-qualitative observations, range limitation, SMOS, ice thickness, TOPAZ4, EnKF-SQ.

*Corresponding author (abhishek.shah@nersc.no)

1 Introduction

Sea ice plays a crucial role in the Arctic climate as it modulates the exchange of heat and moisture between the ocean and the atmosphere (Aagaard and Carmack, 1989; Screen and Simmonds, 2010). Different studies have shown that accurate knowledge of the Sea Ice Thickness (SIT) is beneficial for the Arctic sea ice predictability (Day *et al.*, 2014; Collow *et al.*, 2015; Guemas *et al.*, 2014). The SIT observations from the European Space Agency (ESA) Soil Moisture and Ocean Salinity (SMOS) mission are available in near-real time, at daily frequency during the cold season (October-April). The retrieval method for SMOS SIT observations is based on measurements of the brightness temperature at a low frequency microwave (1.4 GHz, L-band: wavelength of 21 cm) (Kaleschke *et al.*, 2010). The representative depth for the L-Band microwave frequency into the sea ice is about 0.5 m for first-year level ice (Kaleschke *et al.*, 2010; Huntemann *et al.*, 2014). Few studies have shown that assimilating thin SIT from SMOS into coupled ice-ocean model, using ensemble based Data Assimilation (DA) techniques, is able to improve the SIT forecast without being detrimental to other properties (e.g., Yang *et al.*, 2014; Xie *et al.*, 2016; Fritzner *et al.*, 2019). All of these studies, however, ignore the saturated observations of thick ice.

Measurements of thick sea ice on basin-wide scales are also available from laser altimeters onboard ICESat (Forsberg and Skourup, 2005) or from radar altimeters on the European Remote Sensing (ERS), Envisat, CryoSat-2 and Sentinel-3 (Connor *et al.*, 2009; Laxon *et al.*, 2013; Ricker *et al.*, 2014). CryoSat-2 SIT is provided in near-real time (Tilling *et al.*, 2016) but still contains considerable large uncertainties caused by the lack of auxiliary data on snow depth. These uncertainties are proportionally larger for thin ice (i.e., < 1 m) and hence CryoSat-2 practically measures thick sea ice only. A merged product of weekly SIT observations in the Arctic from the CryoSat-2 altimeter and SMOS radiometer, referred as CS2SMOS, has also been developed by combining the two complementary datasets (Kaleschke *et al.*, 2015; Ricker *et al.*, 2017) and made available during the winter months since October 2010. However, the combination of the two satellites is not perfect as biases have been revealed on overlapping areas (Wang *et al.*, 2016; Ricker *et al.*, 2017). Recently, Xie *et al.* (2018) successfully assimilated the merged SIT product CS2SMOS into the TOPAZ4 coupled ocean-sea ice reanalysis system (Sakov *et al.*, 2012) for the Arctic.

While assimilating a merged SIT map, rather than two satellites data streams is practically convenient, the uncertainty of the merged data is more difficult to quantify and bad quantification of the uncertainty may affect the assimilation performance negatively (Mu *et al.*, 2018)[i]. The ability to use well-justified observation errors in data assimilation is sufficiently important to motivate the assimilation of the two separate SIT data streams rather than one merged product. This implies that their detection limits should be taken into account by the data assimilation

[i]It should be noted that the comparison of assimilating merged versus separate data is not informative because their observation errors are not equivalent

method.

In DA, observations are used to reduce the error of the state variables so that the forecast skill can be enhanced. Many observations can only be retrieved within a limited interval of the values that the observed quantity would take in nature. In other words, observations may have a detection limit. One such example is the aforementioned observation of SIT from SMOS. Although, the SIT observations with detection limit do not provide quantifiable data (hard data) above its detection limit, they do give qualitative information (soft data). For instance, the ice could be thicker than a known threshold. Studies from Shah *et al.* (2018) and Borup *et al.* (2015) have shown that assimilating soft data with linear and non-linear toy models using ensemble-based DA methods have the potential to improve the accuracy of the forecast. Therefore, not considering soft data in the assimilation procedure is a potential loss of meaningful information.

Assimilating only thin ice observations, as in Xie *et al.* (2016, Figure 5 and 6), induces a low bias, which is caused by the partial nature of the observation of thin ice. With a new method intended for semi-qualitative data as the EnKF-SQ, the question arise whether this bias can be mitigated or not? The comparison of the EnKF-SQ to the perfect Bayesian solution (Shah *et al.*, 2018) shows that the EnKF-SQ analysis does not coincide with the Bayesian posterior and bears inherent biases: in the case of hard data, the Bayesian and EnKF-SQ posteriors are nearly the same. However, for out-of-range observations and mode of a prior within the observable range, only the maximum likelihood of the EnKF-SQ analysis is preserved but its distribution is flatter than the Bayesian solution with a thicker tail in the unobservable range, so the expectation is too high. Based on this, the EnKF-SQ is expected to be unbiased for thin SIT observations. Nevertheless, it should show a positive bias for out-of-range observations. Further, the thicker tail of the EnKF-SQ analysis distribution in the unobservable range makes it relatively skewed, which is undesirable in a Kalman filtering context.

In this study, we implement and test the overall performance of the stochastic ensemble Kalman filter semi-qualitative (EnKF-SQ) (Shah *et al.*, 2018) in a twin experiment where synthetic SMOS-like SIT observations, with an upper detection limit, are assimilated into a coupled ocean-sea ice forecasting system. The objective is to test the potential of the EnKF-SQ for assimilating soft data with a state of the art ocean and sea ice prediction system, namely TOPAZ4. In addition, a number of single-cycle assimilation experiments using the EnKF-SQ are performed to investigate the sensitivity to the ensemble size and out-of-range observation uncertainty.

This paper is organized as follows: Section 2 introduces the main components of the TOPAZ4 system including the model and the EnKF-SQ DA scheme used in the assimilation experiments. In Section 3, the synthetic ice thickness data are outlined together with the assimilation setup. Section 4 discusses the results of the various assimilation experiments. A general discussion of the study concludes the paper in Section 5.

2 The TOPAZ system

2.1 Model setup

The ocean general circulation model used in the TOPAZ4 system is the version 2.2 of the Hybrid Coordinate Ocean Model (HYCOM) developed at the University of Miami (Bleck, 2002; Chassignet *et al.*, 2003). The TOPAZ4 implementation of HYCOM uses hybrid coordinates in the vertical, which smoothly shift from isopycnal layers in the stratified open ocean to z level coordinates in the unstratified surface mixed layer.

The HYCOM ocean model is coupled to a one-thickness category sea ice model. The single ice thickness category thermodynamics are described in Drange and Simonsen (1996) and the ice dynamics use the Elastic-Viscous-Plastic (EVP) rheology of Hunke and Dukowicz (1997) with a modification from Bouillon *et al.* (2013). The momentum exchange between the ice and the ocean is given by quadratic drag formulas. The model has a minimum thickness of 10 cm for both new and melting ice.

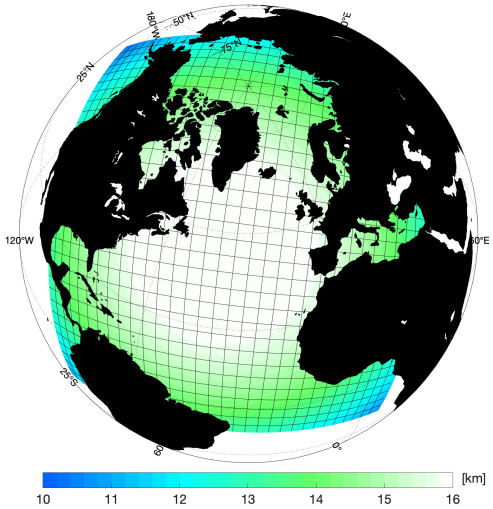


Figure 1: The TOPAZ4 model domain. Background color shading shows the horizontal grid resolution (km) while solid black color represents land.

The model domain covers the North Atlantic and Arctic basins as shown in Figure 1. The model grid is created with conformal mapping (Bentsen *et al.*, 1999) and has a quasi-homogeneous horizontal resolution between 12 − 16 km in the whole domain. The grid has 880×800 horizontal grid points.

2.2 The Ensemble Kalman Filter Semi-Qualitative, EnKF-SQ

The EnKF-SQ (Shah *et al.*, 2018) uses an ensemble of model states to estimate the error statistics closely following the stochastic EnKF algorithm (Burgers *et al.*, 1998; Evensen, 2004). The stochastic EnKF is a two-step filtering method alternating forecast and analysis steps. In the forecast step, the ensemble of model states is integrated forward in time and when observations become available, an analysis of every forecast member, $\mathbf{x}_i^f$ for $i \in 1,2,...,N$, is computed as follows:

$$\mathbf{x}_i^a = \mathbf{x}_i^f + \mathbf{K}(\mathbf{y}_i - \mathbf{H}\mathbf{x}_i^f), \tag{1}$$

$$\mathbf{K} = \mathbf{P}^f \mathbf{H}^T (\mathbf{H}\mathbf{P}^f \mathbf{H}^T + \mathbf{R})^{-1}, \tag{2}$$

where $\mathbf{K}$ is the Kalman gain matrix; $\mathbf{x}_i$ is the i^{th} ensemble state member; $\mathbf{H}$ is the observation operator, mapping the state variable to the observation space (could be non-linear); $\mathbf{R}$ is the observation error covariance matrix; $\mathbf{y}_i$ is the i^{th} perturbed observation vector generated from $\mathcal{N}(\mathbf{y},\mathbf{R})$ and $\mathbf{P}^f$ is the ensemble forecast error covariance matrix. The superscripts a, f, and T stand for analysis, forecast, and matrix transpose, respectively. In practice, $\mathbf{P}^f$ is never computed explicitly and is instead decomposed as follows:

$$\mathbf{P}^f = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{x}_i^f - \bar{\mathbf{x}}^f)(\mathbf{x}_i^f - \bar{\mathbf{x}}^f)^T, \tag{3}$$

where $\bar{\mathbf{x}}^f$ is the mean of the forecast ensemble.

The EnKF-SQ is intended to explicitly assimilate observations with a detection limit. These are divided into two categories depending on whether they are within or outside the observable range. If the observed quantity is within it, the quantitative (hard) data is assimilated as in the stochastic EnKF, otherwise it is considered a qualitative (soft) data and treated differently.

The specific value and error statistics of the out-of-range (OR) observations are unknown. In order to assimilate OR observations, an assumption needs to be made about its likelihood. Following Shah *et al.* (2018), a virtual observation is created at the detection limit and then a two-piece Gaussian observation likelihood is constructed around it. A two-piece Gaussian distribution is obtained by merging two opposite halves of two different Gaussian probability density functions (pdfs) at their common mode, given as follows:

$$f(x) = \begin{cases} we^{-(x-\mu)^2/2\sigma_{ir}^2}, & x \leq \mu, \\ we^{-(x-\mu)^2/2\sigma_{or}^2}, & x > \mu, \end{cases} \tag{4}$$

where $w = \sqrt{\frac{2}{\pi}}(\sigma_{ir} + \sigma_{or})^{-1}$ is a normalizing constant, μ is the detection limit and also the common mode of two different normal distribution; σ_{ir} and σ_{or} are in-range and OR observation error standard deviations (std), respectively.

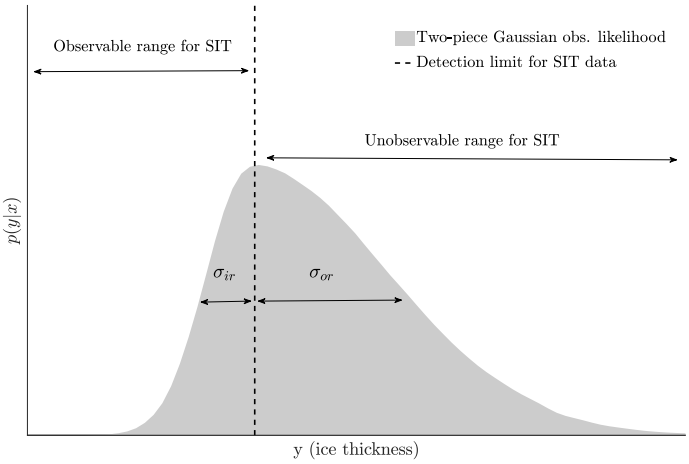


Figure 2: Illustration of the two-piece Gaussian OR-observation likelihood for SMOS-like thin SIT. σ_{ir} is an in-range and σ_{or} is the out-of-range observation error standard deviations, respectively.

Figure 2 is an illustration of a two-piece Gaussian observation likelihood for OR SIT observations. On the left hand side of the detection limit, it is assumed that σ_{ir}, inside the observable range, is defined by the observation uncertainty of hard data at the detection limit. An observation could possibly fall outside the detection limit, due to observation errors, even though its true value is within the observable range. On the right hand side, it is assumed that the σ_{or} (eq. 5) in the unobservable range is defined with the help of a climatological mean for SIT above the detection limit so that extremely high values, which are usually less realistic, receive a lower likelihood (Shah *et al.*, 2018).

$$\sigma_{or} = \underbrace{\int_{\mu}^{+\infty} y f_c(y) dy}_{\text{Climatological mean}} - \mu. \tag{5}$$

$f_c(y)$ is the pdf of the climatological data of the observed quantity. The two-piece Gaussian observation likelihood for soft data is denoted, hereafter, by $\mathcal{N}_{2p}(\mu, \sigma_{ir}^2, \sigma_{or}^2)$. The EnKF-SQ pre-processes the observations by sorting them as either hard y^h or soft y^s. The observation errors are assumed uncorrelated in space, i.e. $\mathbf{R}$ is diagonal.

Update step of the EnKF-SQ

For each forecast member $\mathbf{x}_i^f$ $(i \in 1, 2, \ldots, N)$:

1. For each soft data y_j^s, check whether the observed forecast ensemble member is within

the observable range or not.

2. If $\mathbf{H}_j\mathbf{x}_i^f \leq \mu$, set observation error variance $\mathbf{R}_{j,j} = \sigma_{ir}^2$ otherwise $\mathbf{R}_{j,j} = \sigma_{or}^2$ implying that members inside (outside) the observable range are updated with data parameterized using in-range σ_{ir}^2 (out-of-range σ_{or}^2).

3. After looping over all soft data, compute the Kalman gain $\mathbf{K}_i$ as in Eq. 2 with the updated observation error covariance matrix $\mathbf{R}$. For each $\mathbf{x}_i^f$, a different Kalman gain $\mathbf{K}_i$ is calculated.

4. Evaluate the i^{th} analysis member $\mathbf{x}_i^a$ as in Eq. 1 using $\mathbf{K}_i$. The perturbed observations are generated by sampling from $\mathcal{N}(y_j^h, \sigma_h^2)$ and $\mathcal{N}_{2p}(\mu, \sigma_{ir}^2, \sigma_{or}^2)$[ii] for y_j^h and y_j^s, respectively. σ_h^2 is the observation error variance for y_j^h.

Loop to next member i.

Repeating this process for all forecast members yields the analysis ensemble. For a detailed description of the EnKF-SQ the reader is referred to Section 2 of Shah *et al.* (2018).

3 Experimental Setup

3.1 The synthetic Sea Ice Thickness Data

The synthetic SIT data used in this study is intended to mimic the SIT data from the SMOS mission with an upper detection limit. In order to evaluate the EnKF-SQ method against a perfectly known truth, synthetic observations are generated using the coupled ocean and one-thickness category sea ice model described earlier in Section 2.1. A reference *truth* run (also called nature run) is produced by integrating the coupled ocean sea ice model from 1 January, 2014 to 31 December, 2015 using unperturbed atmospheric forcing from ERA-Interim (Dee *et al.*, 2011). The run is initialized using member number 100 from the 100-member ensemble reanalysis of Xie *et al.* (2017) on 31 December, 2013.

Synthetic SIT data are then generated for the duration of the assimilation experiment from 11 November 2014 to 31 March 2015 by perturbing the truth with Gaussian noise of zero mean and standard deviation σ_{obs}; parameterized as:

$$\sigma_{obs} = 0.06t + 0.05, \tag{6}$$

where t is the truth for ice thickness in meters. The parameterization is chosen such that observation errors increase for thicker ice, which is a general behaviour of positive-valued variables

[ii] σ_{ir} is a special case of σ_h, for hard data at the detection limit.

like SIT. The relationship is obtained through regression of the absolute difference of the daily averaged SIT between the reanalysis product (Xie *et al.*, 2017) and the aforementioned reference trajectory from the month of December 2014 to January 2015. The resulting relationship (not shown here) is linear with a positive slope. SIT observation error represented in Eq. 6 is also qualitatively in line with those used by Xie *et al.* (2016) for SMOS data.

A single upper detection limit of 1 m is imposed on the generated SIT observations, as an analogous for saturation of SMOS data in thick sea ice. The SIT observations are assumed available on every grid cell (except along the coastline) and assimilated on a weekly basis. This is a reasonable assumption as SMOS data comes with a resolution of ($\sim$ 12.5 km), which is also the resolution of the TOPAZ4 system. Model and observation grids are collocated, thus our experiments neglect potential errors due to interpolation, which is out of the scope of this study.

3.2 Out-of-range SIT Climatology

A trivial alternative to the EnKF-SQ in the presence of soft data would be to assimilate climatological data as hard data. It is, therefore, worth investigating how beneficial the assimilation of soft data with the EnKF-SQ is compared to assimilating climatology.

An out-of-range, location-dependent, SIT climatology is computed by taking a time average of the truth (described earlier) for SIT above the detection limit in each grid cell. Averaging is done from January 2014 to December 2015, a period that includes two summers and two winters and encompasses the assimilation period. Even though the latter takes place in winter, the climatology has a high bias because by construction it only contains SIT above 1 m. The observation error variance for the climatological value is also location-dependent, equal to the variance of all reference truth values above the detection limit in the same grid cell.

3.3 Assimilation setup

In contrast to earlier TOPAZ4 studies that updated the whole water column variables (Xie *et al.*, 2018), here the state vector $\mathbf{x}$ consists of only two sea ice variables: SIT and sea ice concentration (SIC). This therefore constitutes a case of a weakly coupled assimilation where the ocean is only updated by dynamical re-adjustments from the sea ice updates. Kimmritz *et al.* (2018), have shown that while strongly coupled ocean and sea ice is clearly beneficial, weakly coupled DA can still achieve reasonable results.

In the analysis, sampling errors in the forecast error covariance can give rise to spurious correlations between remote grid points, a problem which may become more pronounced for smaller ensemble sizes (Houtekamer and Mitchell, 1998). A common practice to counteract sampling errors is to perform local analysis in which variables at each grid cell are updated using only

the observations within a radius of influence r_o around the grid cell (Houtekamer and Mitchell, 1998; Evensen, 2003). For simplicity, a single closest local observation within $r_o = 300$ km is used here during the analysis.

In TOPAZ4, model error is introduced by increasing the model spread via perturbing few forcing fields. The perturbations are pseudo-random fields computed in a Fourier space with a decorrelation time-scale of 2 days and horizontal decorrelation length scale of 250 km, as described in Evensen (2003). Perturbed variables include air temperature, wind speeds, cloud cover, sea level pressure (Sakov *et al.*, 2012, Section 3.3) and yield curve eccentricity in the EVP rheology (Hunke and Dukowicz, 1997, Table 1). In addition, precipitation is also perturbed with log-normal noise and standard deviation of 100%. This affects the snowfall when temperatures are below zero. Snow is an important thermal insulator and therefore hampers sea ice growth/melt.

3.4 Target Benchmarks

The performance of the EnKF-SQ is compared against three different versions of the stochastic EnKF and a Free run, denoted as follows:

1. **EnKF-ALL**: No detection limit is applied on SIT observations thus even thick ice data from the reference run is assimilated. This run acts as an upper bound for performance because it is the only one that assimilates out-of-range observations as hard data with known statistics, which can be seen as *cheating*.

2. **EnKF-CLIM**: The SIT climatology with climatological variance is assimilated instead of hard data.

3. **EnKF-IG**: Only hard data is assimilated and soft data is ignored, similar to Xie *et al.* (2016). This run is meant to assess the added value of the EnKF-SQ.

4. **Free-run**: The Free-run is the average of the 99 members without DA. It is run with perturbations, contrarily to the aforementioned single-member truth run.

To evaluate the performance of the different DA methods, we compute the root mean square error (RMSE) of the ensemble mean at time t as:

$$\mathrm{RMSE}_t = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\bar{\mathbf{x}}^f_{i,t} - \mathbf{x}^r_{i,t}\right)^2}, \tag{7}$$

where $\mathbf{x}^r$ and $\bar{\mathbf{x}}^f$ is the n-dimensional reference (unperturbed truth) and mean of the prior state vector at time t, respectively. We also monitor the average ensemble spread (AES) for

each filter, which we calculate at every assimilation cycle as:

$$\mathrm{AES}_t = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\sigma_{i,t}^2}, \tag{8}$$

where $\sigma_{i,t}^2$ can either be the prior or posterior ensemble variance at time t, respectively.

3.5 Ensemble size

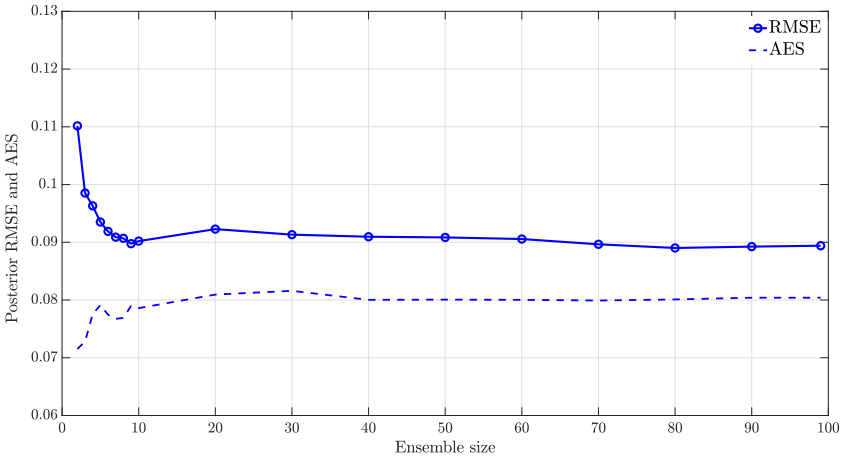


Figure 3: Time-averaged posterior RMSE and AES resulting from single cycle assimilation runs for different ensemble sizes using EnKF-SQ.

In order to select the ensemble size, single-cycle assimilation sensitivity experiments are conducted using EnKF-SQ by varying the ensemble size between 2 and 99. The resulting time-averaged RMSE and AES of the posterior SIT estimates are displayed in Figure 3. The plot indicates that for $N \geq 10$, there is no significant difference in the performance of the EnKF-SQ. This is mostly due to the small size of the local state vector; consisting only of two variables. An ensemble as small as 10 members is however less likely to succeed on the long term especially if the number of state variable and observations increase. Results from the other three EnKF runs (not shown) showed the exact same behavior. Thus, the initial ensemble is set as the first 99 members of the reanalysis ensemble of Xie *et al.* (2016) on 31 December 2013. The initial ensemble is then spun up from January, 2014 until the start of the assimilation experiment (i.e., November 11) with perturbed forcing to increase the variability. As described earlier, member number 100 of the reanalysis run was used to generate the truth in this study.

The assimilation framework is sub-optimal for few reasons, in particular because of the weakly coupled updates. Further, SIT errors are erroneously assumed Gaussian while they are not.

These sub-optimalities are not uncommon in realistic applications. They do cause some limited loss of performance but generally do not prevent us from applying the EnKF.

In terms of computational resources, we used a single processor on supercomputer for each of the four DA methods. The total wall-clock time required by each analysis scheme, to update the SIT and SIC state variables along with the IO operations, is approximately 6 minutes on a 1.4GHz Cray XE6. This is much less than the TOPAZ4 one-week forward model run, for which each member runs on 134 parallel processors in approximately 5 minutes.

4 Assimilation Results

4.1 Tuning the EnKF-SQ out-of-range likelihood

The out-of-range standard deviation σ_{or} is the only new parameter introduced into the EnKF-SQ compared to the stochastic EnKF. Therefore, it is important to study how the uncertainty in the estimate of σ_{or} affects the performance of the EnKF-SQ scheme. For this, we carried out a number of single-cycle assimilation experiments by introducing a scalar multiplier α to equation 5 such that $\sigma^*_{or} = \alpha \cdot \sigma_{or}$.

RMSE and AES of the posterior SIT estimates are plotted in Figure 4 for a wide range of α, varying between 0.1 and 3.0. Such a range is very broad for most realistic applications. $\alpha < 0.4$ strongly degrades the accuracy of the EnKF-SQ along with significant decrease in the AES. The large difference between RMSE and AES values, indicate a possible filter divergence. This is because for small α values, the sampling of a two-piece Gaussian likelihood for observation perturbations is prone to generate samples concentrated around the detection limit, thus pulling the analysis close to the detection limit, subsequently reducing the ensemble spread and increasing the RMSE. As α approaches 1, the RMSE attains the minimum value and further becomes consistent with the AES. When α increases beyond 2, the sampling of OR likelihood starts producing large perturbations, which makes the analysis increment capricious and eventually deteriorates the performance of the EnKF-SQ. Accordingly, in what follows we set $\alpha = 1$.

To illustrate how the EnKF-SQ updates the SIT by assimilating range-limited SIT observations, we plot the prior mean (Figure 5a) and analysis increment (Figure 5b) on 11 November 2014. The solid black line on both maps is the isoline for 1 m of SIT. The forecast places the thick ice (up to 3 m) north of Greenland and north-eastern part of Canada. The increments are not only visible outside of the 1 m isoline but also inside the central Arctic region where only soft data are assimilated. It is important to notice that there is nearly zero increment in the central Arctic region and the Beaufort sea where the sea ice is thicker than 1.5 m. This is because the EnKF-SQ analysis do not impose strong updates on the prior if it is above the detection limit

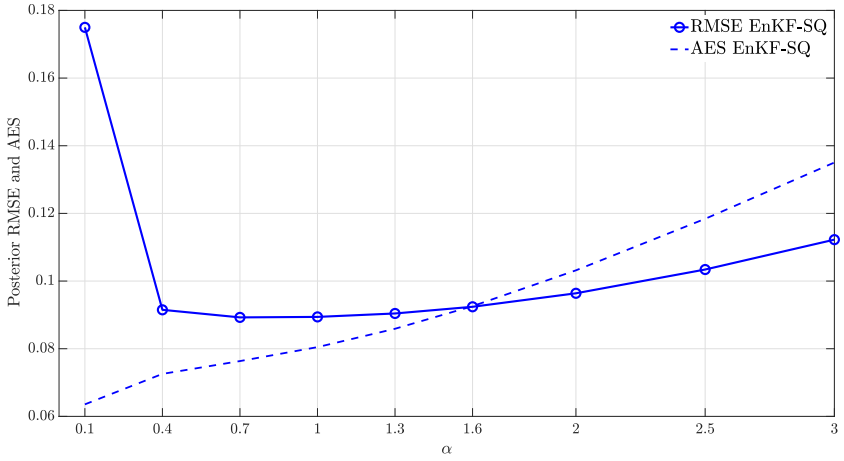


Figure 4: Time-averaged posterior RMSE and AES resulting from single cycle assimilation runs for a wide range of the multiplicative factor α.

and observations are out-of-range.

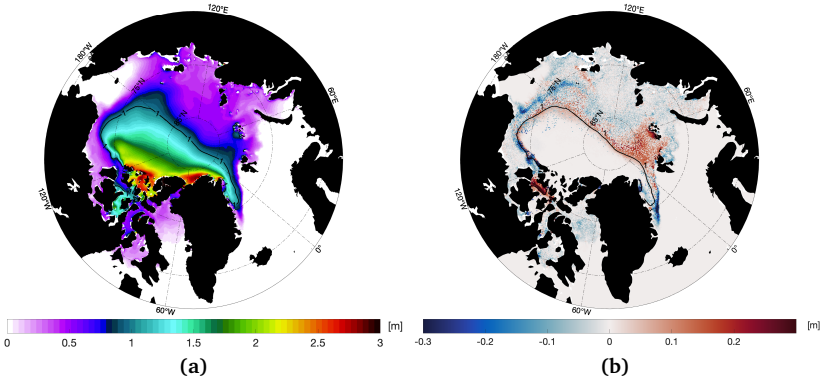


Figure 5: (a) Prior ensemble mean of the ice thickness on 11 November 2014. The solid black line is the 1 m SIT isoline. (b) The increment (analysis-forecast) for SIT after incorporating the observations.

4.2 Performance Assessment

Figure 6 shows the time evolution of the RMSE and AES of the prior SIT estimates obtained using the EnKF-ALL, EnKF-SQ, EnKF-CLIM, EnKF-IG and the Free-run. The percentage of OR observations (to the total number of observations) available at every cycle is added to the plot. As expected, EnKF-ALL outperforms all other schemes while EnKF-IG is the least accurate. It should be noted that there is an increasing trend in the RMSE, which is seasonally driven; a

similar behavior reported in Xie *et al.* (2016). Assimilating soft data with the EnKF-SQ clearly improves the prior RMSE compared to the EnKF-IG. This is consistent over the entire assimilation period. The number of OR observations gradually increases as the cold season intensifies leaving only a few hard data during the months of February and March 2015. Even with a very limited number of hard data, the EnKF-SQ outperforms EnKF-IG. The RMSE resulting from the EnKF-CLIM is marginally higher than that of the EnKF-SQ, except during the last three months of the assimilation experiment. The reason for this could be twofold: (i) In the early stages of the experiment, the climatology tends to overestimate SIT due to the large seasonal cycle compared to later months. This causes the climatology to pull the update towards large values and hence degrades the performance of the EnKF-CLIM. (ii) Fewer hard data leads to larger RMSE values in the EnKF-SQ as can be seen towards the end of winter and start of the spring. Overall, the RMSE and AES show consistent ensemble statistics such that sufficient variability is preserved in the system after cycling over time.

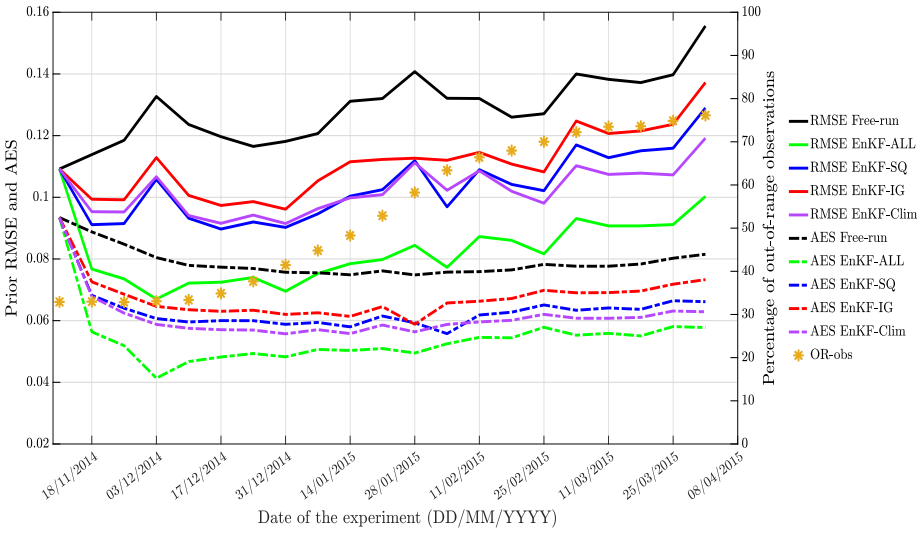


Figure 6: Left y-axis: Time evolution of the prior RMSE (solid lines) and AES (dashed lines) for SIT estimates. Right y-axis: The orange asterisks represent the percentage of the out-of-range observations during assimilation resulting from the EnKF-SQ, EnKF-CLIM and EnKF-IG.

In order to visualize area-wise improvements, we plot the map of time-averaged RMSE of the SIT prior estimates in Figure 7. The EnKF-ALL yields the best RMSE throughout the entire region. Compared to the EnKF-IG, the EnKF-SQ performs better in the central Arctic region, Greenland's north-eastern shelf, the Canadian Arctic Archipelago and in the Beaufort Sea. On average, the EnKF-SQ and EnKF-CLIM estimates are approximately 8% more accurate than those of the EnKF-IG.

The EnKF-CLIM, seems to produce larger improvements than the EnKF-SQ specifically along the

Ellesmere island. However, it also increases the prediction error in the Beaufort sea more than that of the EnKF-IG. A number of reasons may explain this behavior. The climatology being too high compared to the seasonal mean yields an artificial increase of the model thickness, which happens to agree with the truth along the Ellesmere island. The recurrent update due to the assimilation of climatology is propagated dynamically by the Beaufort gyre into the Beaufort sea creating an anomaly compared to the truth, which is not thicker.

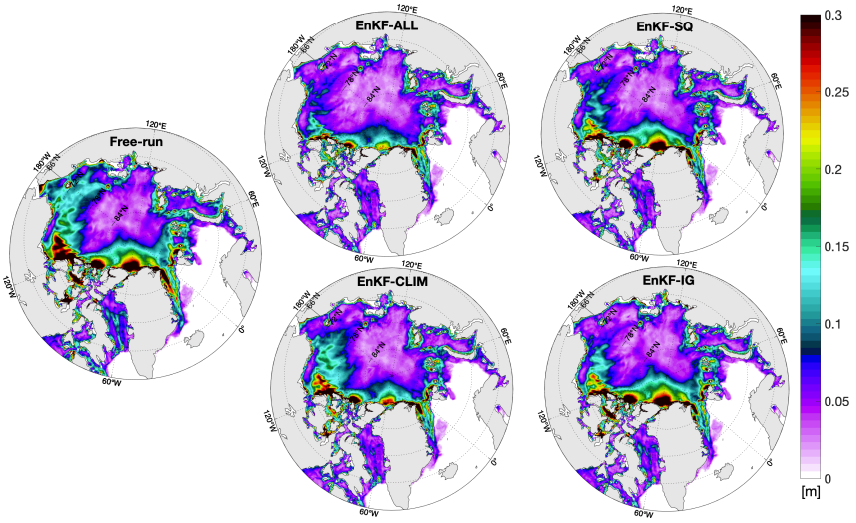


Figure 7: Maps of time-averaged prior RMSE for SIT obtained using: EnKF-ALL (top left), EnKF-SQ (top right), Free-run (center left), EnKF-CLIM (bottom left) and EnKF-IG (bottom right). Averaging is done over the period of experiment, i.e., from November 2014 to March 2015.

The analysis algorithm of the EnKF-SQ is designed such that improvements are expected mostly where SIT is close to the threshold. As a way to examine this, we computed the time-averaged RMSE of the prior SIT estimates for different ice thickness intervals of 25 cm using all DA schemes (Figure 8). The values on the x-axis of Figure 8 represent the upper bounds of each 25 cm SIT bin interval except for the first bin of size 10 cm because of the model 10 cm minimum thickness. The RMSE for all DA schemes within each SIT bin is computed by finding the location of grid cells for which the observations fall within the bin interval.

Figure 8 suggests that RMSE values for all schemes below 1 m of SIT are approximately the same, as they all assimilate hard data. Once SIT increases beyond the detection limit, EnKF-ALL becomes the most accurate followed by the EnKF-SQ up to SIT of 2 m. The EnKF-SQ performs as expected for observation values in the vicinity of the detection limit where the assimilation of soft data is clearly enhancing the accuracy compared to the EnKF-IG and EnKF-CLIM. The performance of the EnKF-SQ is not as good as the EnKF-CLIM for thicker ice, which can also be seen in Figure 7 around the northern coast of Greenland. It is worth noticing that even though there is no data to assimilate for SIT $>$ 1 m in the EnKF-IG scheme, it is performing better than

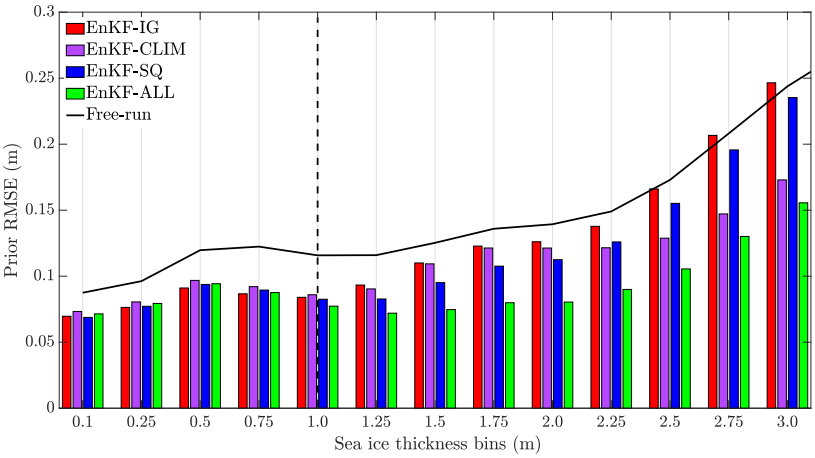


Figure 8: Bar chart of time-averaged conditional prior RMSEs for SIT obtained using all tested DA schemes. Solid black line represents time averaged Free-run RMSE. Black dashed line depicts the 1 m detection limit. The x-axis denotes the SIT bins with bin size of 25 cm. The values on x-axis are the upper bounds of the SIT for that particular bin.

the Free-run up to 2.25 m of SIT. This advantage has been previously reported by Xie *et al.* (2016, see Figure 8) and can be either due to the reduction of the positive bias in the free run (shown in Figure 9) by assimilating thin ice only or due to dynamical model adjustments after assimilation. In other words, improvements to thin ice are propagated in time to the period where ice gets thicker.

4.3 Bias and Skewness Analysis

The EnKF-IG updates the prior members by only assimilating observations of thin ice with a maximum thickness of 1 m. This causes the algorithm to introduce negative conditional bias for thick ice (knowing that the observation is thin ice, the assimilation reduces the ice thickness more that it can thicken it). Similarly, the EnKF-SQ update may introduce a bias towards the detection limit due to assimilation of soft data and the EnKF-CLIM towards the climatology. To investigate these likely biases in different DA schemes, we present a bar chart of time-averaged conditional bias for the posterior estimates of SIT in Figure 9. The conditional bias is calculated by finding the location of the grid cells for which the observations fall within the SIT bin interval. The positive values represent an overestimation of SIT after the assimilation and vice versa.

The four DA runs exhibit a small negligible positive bias of approximately 0.5 to 1 cm for thin ice. The Free-run bias, on the other hand, is larger than ~ 6 cm. Above the threshold limit, there is a clear positive bias of 5 to 7 cm in the EnKF-CLIM posterior estimates, up until 2 m. As

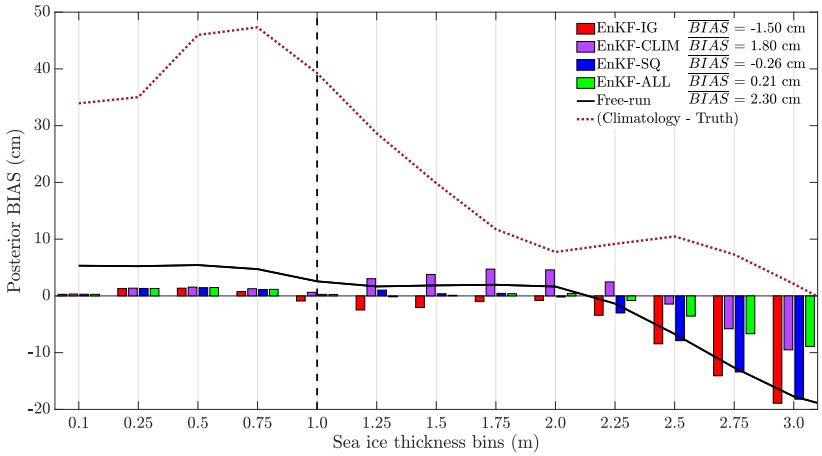


Figure 9: Bar chart of time-averaged posterior bias for SIT obtained from all tested DA schemes. Solid black line represents the time-averaged bias for SIT obtained using the Free-run. Red dotted line represents the time-averaged difference of climatology and truth. SIT bins are displayed on the x-axis with a bin size of 25 cm. Reported in the legend are the time-averaged-weighted total mean bias including the bins for ice thicker than 3 m, which is not shown here. The weights are computed as a fraction of the number of grid cells falling in specific bin interval over total number of grid cells.

seen earlier, the climatology tends to overestimate the truth during the first few months of the experiment when the ice is thin (red dotted line in the Figure 9). EnKF-IG estimates, over the same interval, exhibit a small negative bias, possibly left over from the conditional assimilation of thin ice. It is important to note that there is almost zero bias in the EnKF-SQ estimates, matching that of the EnKF-ALL for $1 \leq \text{SIT} \leq 2$ m.

There is a systematic increasing negative bias for SIT > 2 m, which reaches almost 20 cm for SIT $= 3$ m in the Free-run, EnKF-IG and EnKF-SQ. A similar trend of negative bias is also observed in the EnKF-ALL and EnKF-CLIM runs but to a slightly lesser extent. The negative bias in the Free-Run is likely due to the perturbation of the forcing fields, specifically the wind perturbations, which can cause erratic movements of ice that export thicker sea ice into areas of thinner ice. Since all assimilation runs use perturbed winds, this effect is likely to impact the EnKF-IG and EnKF-SQ more than the EnKF-ALL and EnKF-CLIM. In addition, it is important to mention that there are fewer grid points (not shown here) in the bins for thicker ice compared to thin ice, which may also affect the estimation of the bias for these bins, making them statistically less significant.

As discussed in Shah *et al.* (2018), the two-piece Gaussian observation likelihood may influence the shape of the posterior distribution, making it skewed and thus less Gaussian. In order to examine this, we evaluate and plot the conditional skewness of the posterior estimates of SIT only at the last assimilation step in Figure 10. The conditional skewness of the posterior is calculated as the average value of the skewness for all grid cells where the truth falls within

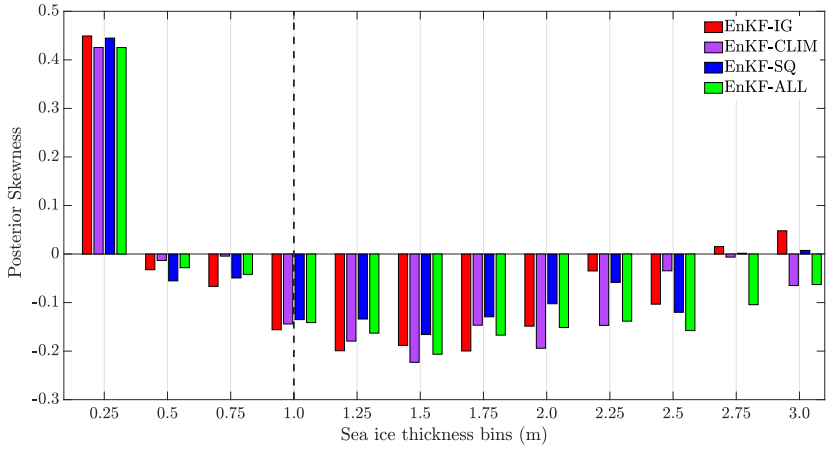


Figure 10: Bar chart of the conditional posterior skewness for SIT estimates obtained using all tested DA schemes and computed at the final assimilation time. Dashed black line represents the detection limit of 1 m SIT.

the interval of a bin in consideration. Note that contrary to the computation of the conditional BIAS at the location of the observations, the conditional skewness is computed at the truth locations.

As shown in the figure, thin ice (SIT ≤ 25 cm) yields noticeable skewness in the posterior estimates for all schemes. In the first bin, the truth is close to zero meters (open water) and hence all instances where thin ice has melted in the assimilation run count as zero value. On the other hand, freezing instances lead to various thickness values above 25 cm. Both effect together can make the distribution skewed. The bin between zero and 10 cm shows even larger skewness and has been removed for a better visual presentation. Other than the first bin, a small negative skewness is observed for all the schemes. One possible explanation is the fast melting of ice, drifting over warm waters; a situation enhanced by the lack of coupling with the ocean in the assimilation. This result confirms that the EnKF-SQ, although it uses a skew 2-piece Gaussian likelihood, does not introduce any noticeable positive skewness in its posterior.

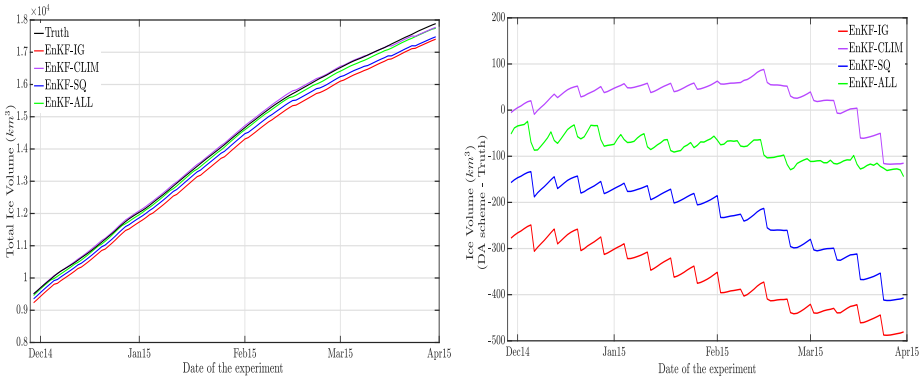
4.4 Physical Consistency

Ice-ocean models are essential tools for computing integrated quantities that are often difficult to estimate from observations only. Sea ice volume and water transport between ocean basins are such high interest quantities for climate studies. Therefore, it is important to evaluate these quantities to verify that the use of data assimilation does not cause physical inconsistencies.

The total sea ice volume is the integral of sea ice concentration times the sea ice thickness over

the entire model area. Its evolution for the different assimilation runs is shown in Figure 11a. The difference between the assimilation runs compared to the true sea ice volume (Figure 11b) is relatively small. This is because none of the DA schemes has extensively added or removed ice during the assimilation run. In Figure 11b a classical seesaw Kalman update behavior is observed. The comparison also reveals that most methods tend to underestimate the ice volume except for EnKF-CLIM.

As described earlier, the EnKF-IG has a negative SIT bias, which translates to a nominal loss of between 300 km^3 to 500 km^3 of sea ice volume from the beginning to the end of the winter (less than 3% of the total simulated ice volume). Seesaw of the time series curves confirm that the EnKF-IG update does remove some ice, which grows back during the subsequent TOPAZ4 model run. The EnKF-SQ does only partially mitigate this loss by 100 to 200 km^3 of ice. Surprisingly, the EnKF-ALL is not bias-free either with a loss of up to 100 km^3 of ice, which can be caused by various sub-optimal aspects of the data assimilation system, in particular the aforementioned effect of wind perturbations on the areas of thickest ice and the weakly coupled DA. These effects also contribute to the low bias in the other two methods.



(a) **(b)**

Figure 11: (a) Daily ensemble average of sea ice volume over the TOPAZ4 model area for the entire experiment time. (b) Difference of sea ice volume from the truth and all tested DA schemes.

The EnKF-CLIM ice volume is closest to the truth run with a little overestimation in the beginning of the winter, then an underestimation in the spring. The construction of the climatological data can explain this trend: since SIT data above one meter only have been retained, the climatology overestimates the SIT in the beginning of the winter but then underestimates the SIT in the midst of the winter because it also accounts for summer SIT. A different construction of the SIT climatology data would have led to a different tendency in EnKF-CLIM.

5 Discussion and Conclusions

The purpose of this paper is to demonstrate the usefulness of assimilating range-limited observations with the new EnKF-SQ DA scheme under a realistic experimental setup. Compared to the stochastic EnKF, the main algorithmic difference is the need to compute a different Kalman gain for each ensemble member, depending on the location of the member to the threshold when the observation is out-of-range. This does not make the EnKF-SQ less efficient, but rather prevents the algorithm from being included as a simple extension of existing EnKF codes: it cannot be expressed with an ensemble transform matrix.

The assimilation of synthetic sea ice thickness data with a upper detection limit of 1 m in a coupled ice-ocean model of TOPAZ4 is demonstrated using the EnKF-SQ and shown to have a useful impact on SIT estimates. The results obtained with SMOS-like observations can be generalized to CryoSat2-like observations by reversing the upper limit into a lower limit. Thus, merging the two products may not be necessary because each satellite data can be assimilated in a separate EnKF-SQ step.

Different assimilation experiments are conducted to assess the performance of the EnKF-SQ against other EnKF configurations assimilating only thin ice; both thin and thick ice; and climatology during a winter period in the Arctic. The study shows that assimilating soft data improves the forecast accuracy compared to ignoring them by approximately 8%, particularly where sea ice approaches the detection limit. Such a difference can be important in the performance of an operational system.

The performance exhibited by assimilating a reasonably accurate climatology was similar to the EnKF-SQ. Also, our choice of climatology being annual rather than seasonal may explain some of the flaws in the EnKF-CLIM. Nonetheless, the context of twin experiments is very favorable to EnKF-CLIM because the climatological truth is perfectly known; a case which is not true in realistic situations. For instance, in summer there are very few ice thickness measurements and thus it is difficult to construct a meaningful climatology. To this end, it is essential to investigate and compare the performance of the EnKF-SQ and EnKF-CLIM in a context of a biased model twin experiment and with a range of toy models (from linear to non linear regimes).

Assessing the bias of the analysis showed that there is no introduction of any significant bias by the EnKF-SQ, other than the negative bias for thicker ice which is observed in all tested DA schemes. Likewise, the posterior distributions resulting from the application of the EnKF-SQ did not consist of any noticeable higher order moments that could result in undesirable non-Gaussian features because of the two-piece Gaussian likelihood. This is most likely the case for all realistic applications where one would expect relatively small assimilation updates coming regularly in time. Furthermore, the choice of out-of-range (OR) observation error variance was not found to be very critical. A wide range of values for this parameter were tested and lead to acceptable performance of the EnKF-SQ. Ways of estimating σ_{or}^2 adaptively in space and time is

currently being investigated and will be reported in a follow-up study. Concerning the physical constraints of the model, the EnKF-SQ estimates were found to be physically consistent and comparable to other tested assimilation schemes.

The EnKF-SQ therefore makes a viable data assimilation strategy for range-limited observations in high-dimensional nonlinear systems. Future research will focus on assimilating real data, in which the EnKF-SQ is confronted with large observation biases unlike the presented twin experiments setup.

Acknowledgements

The research is funded by the Nordic Center of Excellence EmblA (Ensemble-based data assimilation for environmental monitoring and prediction) under NordForsk contract number 56801. The Copernicus Marine Services and the Nansen Scientific Society have also contributed to the funding. Norwegian grants of the computer time (nn2993k) and data storage space (ns2993k) have also been used for the simulations.

Bibliography

Aagaard K, Carmack EC. 1989. The role of sea ice and other fresh water in the Arctic circulation. *Journal of Geophysical Research: Oceans* **94**(C10): 14 485–14 498, doi:10.1029/JC094iC10p14485, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JC094iC10p14485.

Bentsen M, Evensen G, Drange H, Jenkins AD. 1999. Coordinate transformation on a sphere using conformal mapping. *Monthly Weather Review* **127**(12): 2733–2740, doi:10.1175/1520-0493(1999)127<2733:CTOASU>2.0.CO;2, URL https://doi.org/10.1175/1520-0493(1999)127<2733:CTOASU>2.0.CO;2.

Bleck R. 2002. An oceanic general circulation model framed in hybrid isopycnic-cartesian coordinates. *Ocean modelling* **4**(1): 55–88.

Borup M, Grum M, Madsen H, Mikkelsen PS. 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, URL http://dx.doi.org/10.1007/s00477-014-0908-1.

Bouillon S, Fichefet T, Legat V, Madec G. 2013. The elastic–viscous–plastic method revisited. *Ocean Modelling* **71**: 2 – 12, doi:https://doi.org/10.1016/j.ocemod.2013.05.013, URL

http://www.sciencedirect.com/science/article/pii/S146350031300098X. Arctic Ocean.

Burgers G, Jan van Leeuwen P, Evensen G. 1998. Analysis scheme in the ensemble Kalman filter. *Monthly weather review* **126**(6): 1719–1724.

Chassignet EP, Smith LT, Halliwell GR, Bleck R. 2003. North Atlantic simulations with the hybrid coordinate ocean model (HYCOM): Impact of the vertical coordinate choice, reference pressure, and thermobaricity. *Journal of Physical Oceanography* **33**(12): 2504–2526.

Collow TW, Wang W, Kumar A, Zhang J. 2015. Improving Arctic sea ice prediction using PIOMAS initial sea ice thickness in a coupled Ocean-Atmosphere model. *Monthly Weather Review* **143**(11): 4618–4630, doi:10.1175/MWR-D-15-0097.1, URL https://doi.org/10.1175/MWR-D-15-0097.1.

Connor LN, Laxon SW, Ridout AL, Krabill WB, McAdoo DC. 2009. Comparison of Envisat radar and airborne laser altimeter measurements over Arctic sea ice. *Remote Sensing of Environment* **113**(3): 563 – 570, doi:https://doi.org/10.1016/j.rse.2008.10.015, URL http://www.sciencedirect.com/science/article/pii/S0034425708003283.

Day JJ, Hawkins E, Tietsche S. 2014. Will Arctic sea ice thickness initialization improve seasonal forecast skill? *Geophysical Research Letters* **41**(21): 7566–7575, doi:10.1002/2014GL061694, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2014GL061694.

Dee DP, Uppala SM, Simmons A, Berrisford P, Poli P, Kobayashi S, Andrae U, Balmaseda M, Balsamo G, Bauer dP, *et al.* 2011. The era-interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the royal meteorological society* **137**(656): 553–597.

Drange H, Simonsen K. 1996. Formulation of air-sea fluxes in the ESOP2 version of MICOM. Technical 125, NERSC.

Evensen G. 2003. The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean dynamics* **53**(4): 343–367.

Evensen G. 2004. Sampling strategies and square root analysis schemes for the EnKF. *Ocean dynamics* **54**(6): 539–560.

Forsberg R, Skourup H. 2005. Arctic Ocean gravity, geoid and sea-ice freeboard heights from ICESat and GRACE. *Geophysical Research Letters* **32**(21), doi:10.1029/2005GL023711, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2005GL023711.

Fritzner S, Graversen R, Christensen KH, Rostosky P, Wang K. 2019. Impact of assimilating sea ice concentration, sea ice thickness and snow depth in a coupled ocean-sea ice modelling

system. *The Cryosphere* **13**(2): 491–509, doi:10.5194/tc-13-491-2019, URL https://www.the-cryosphere.net/13/491/2019/.

Guemas V, Blanchard-Wrigglesworth E, Chevallier M, Day JJ, Déqué M, Doblas-Reyes FJ, Fučkar NS, Germe A, Hawkins E, Keeley S, Koenigk T, Salas y Mélia D, Tietsche S. 2014. A review on Arctic sea-ice predictability and prediction on seasonal to decadal time-scales. *Quarterly Journal of the Royal Meteorological Society* **142**(695): 546–561, doi:10.1002/qj.2401, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.2401.

Houtekamer PL, Mitchell HL. 1998. Data assimilation using an ensemble Kalman filter technique. *Monthly Weather Review* **126**(3): 796–811, doi:10.1175/1520-0493(1998)126<0796:DAUAEK>2.0.CO;2, URL https://doi.org/10.1175/1520-0493(1998)126<0796:DAUAEK>2.0.CO;2.

Hunke EC, Dukowicz JK. 1997. An elastic–viscous–plastic model for sea ice dynamics. *Journal of Physical Oceanography* **27**(9): 1849–1867, doi:10.1175/1520-0485(1997)027<1849:AEVPMF>2.0.CO;2, URL https://doi.org/10.1175/1520-0485(1997)027<1849:AEVPMF>2.0.CO;2.

Huntemann M, Heygster G, Kaleschke L, Krumpen T, Mäkynen M, Drusch M. 2014. Empirical sea ice thickness retrieval during the freeze-up period from SMOS high incident angle observations. *The Cryosphere* **8**(2): 439–451, doi:10.5194/tc-8-439-2014, URL https://www.the-cryosphere.net/8/439/2014/.

Kaleschke L, Maaß N, Haas C, Hendricks S, Heygster G, Tonbøe RT. 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**(4): 583–592, doi:10.5194/tc-4-583-2010, URL https://www.the-cryosphere.net/4/583/2010/.

Kaleschke L, Tian-Kunze X, Maas N, Ricker R, Hendricks S, Drusch M. 2015. Improved retrieval of sea ice thickness from SMOS and CryoSat-2. *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)* : 5232–5235.

Kimmritz M, Counillon F, Bitz C, Massonnet F, Bethke I, Gao Y. 2018. Optimising assimilation of sea ice concentration in an Earth system model with a multicategory sea ice model. *Tellus A: Dynamic Meteorology and Oceanography* **70**(1): 1–23, doi:10.1080/16000870.2018.1435945, URL https://doi.org/10.1080/16000870.2018.1435945.

Laxon SW, Giles KA, Ridout AL, Wingham DJ, Willatt R, Cullen R, Kwok R, Schweiger A, Zhang J, Haas C, Hendricks S, Krishfield R, Kurtz N, Farrell S, Davidson M. 2013. CryoSat-2 estimates of Arctic sea ice thickness and volume. *Geophysical Research Letters* **40**(4): 732–737, doi:10.1002/grl.50193, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/grl.50193.

Mu L, Losch M, Yang Q, Ricker R, Losa SN, Nerger L. 2018. Arctic-wide sea ice thickness estimates from combining satellite remote sensing data and a dynamic ice-ocean model with data assimilation during the CryoSat-2 period. *Journal of Geophysical Research: Oceans* **123**(11): 7763–7780, doi:10.1029/2018JC014316, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018JC014316.

Ricker R, Hendricks S, Helm V, Skourup H, Davidson M. 2014. Sensitivity of CryoSat-2 Arctic sea-ice freeboard and thickness on radar-waveform interpretation. *The Cryosphere* **8**(4): 1607–1622, doi:10.5194/tc-8-1607-2014, URL https://www.the-cryosphere.net/8/1607/2014/.

Ricker R, Hendricks S, Kaleschke L, Tian-Kunze X, King J, Haas C. 2017. A weekly Arctic sea-ice thickness data record from merged CryoSat-2 and SMOS satellite data. *The Cryosphere* **11**(4): 1607–1623, doi:10.5194/tc-11-1607-2017, URL https://www.the-cryosphere.net/11/1607/2017/.

Sakov P, Counillon F, Bertino L, Lisæter K, Oke P, Korablev A. 2012. TOPAZ4: an ocean-sea ice data assimilation system for the north Atlantic and Arctic. *Ocean Science* **8**(4): 633.

Screen J, Simmonds I. 2010. The central role of diminishing sea ice in recent Arctic temperature amplification. *Nature* **464**: 1334–7, doi:http://dx.doi.org/10.1038/nature09051, URL http://dx.doi.org/10.1038/nature09051.

Shah A, Gharamti ME, Bertino L. 2018. Assimilation of semi-qualitative observations with a stochastic ensemble kalman filter. *Quarterly Journal of the Royal Meteorological Society* **0**(0), doi:10.1002/qj.3381, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3381.

Tilling RL, Ridout A, Shepherd A. 2016. Near-real-time Arctic sea ice thickness and volume from CryoSat-2. *The Cryosphere* **10**(5): 2003–2012, doi:10.5194/tc-10-2003-2016, URL https://www.the-cryosphere.net/10/2003/2016/.

Wang X, Key J, Kwok R, Zhang J. 2016. Comparison of Arctic sea ice thickness from satellites, aircraft, and PIOMAS data. *Remote Sensing* **8**(9), doi:10.3390/rs8090713, URL http://www.mdpi.com/2072-4292/8/9/713.

Xie J, Bertino L, Counillon F, Lisæter KA, Sakov P. 2017. Quality assessment of the TOPAZ4 reanalysis in the Arctic over the period 1991–2013. *Ocean Science* **13**(1): 123–144, doi:10.5194/os-13-123-2017, URL https://www.ocean-sci.net/13/123/2017/.

Xie J, Counillon F, Bertino L. 2018. Impact of assimilating a merged sea ice thickness from CryoSat-2 and SMOS in the Arctic reanalysis. *The Cryosphere Discussions* **2018**: 1–44, doi:10.5194/tc-2018-101, URL https://www.the-cryosphere-discuss.net/tc-2018-101/.

Xie J, Counillon F, Bertino L, Tian-Kunze X, Kaleschke L. 2016. Benefits of assimilating thin sea ice thickness from SMOS into the TOPAZ system. *The Cryosphere* **10**(6): 2745–2761, doi: 10.5194/tc-10-2745-2016, URL https://www.the-cryosphere.net/10/2745/2016/.

Yang Q, Losa SN, Losch M, Tian-Kunze X, Nerger L, Liu J, Kaleschke L, Zhang Z. 2014. Assimilating SMOS sea ice thickness into a coupled ice-ocean model using a local SEIK filter. *Journal of Geophysical Research: Oceans* **119**(10): 6680–6692, doi:10.1002/2014JC009963, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2014JC009963.

Graphic design: Communication Division, UiB / Print: Skipnes Kommunikasjon AS

uib.no

ISBN: 9788230856741 (print)
9788230863367 (PDF)