Technical note: Sequential ensemble data assimilation in convergent and divergent systems

Hannes Helmut Bauser[1,2], Daniel Berg[2,3], and Kurt Roth[2,4]

[1]Biosphere 2, University of Arizona, USA
[2]Institute for Environmental Physics (IUP), Heidelberg University, Germany
[3]HGS MathComp, Heidelberg University, Germany
[4]Interdisciplinary Center for Scientific Computing (IWR), Heidelberg University, Germany

Correspondence: Hannes Helmut Bauser (hbauser@arizona.edu)

Abstract. Data assimilation methods are used throughout the geosciences to combine information from uncertain models and uncertain measurement data. However, the characteristics of geophysical systems differ and may be distinguished between divergent and convergent systems. In divergent systems initially nearby states will drift apart, while they will coalesce in convergent systems. This difference has implications for the application of sequential ensemble data assimilation methods. This study explores these implications on two exemplary systems: the divergent Lorenz-96 model and the convergent description of soil water movement by the Richards equation. The results show that sequential ensemble data assimilation methods require a sufficient divergent component. This makes the transfer of the methods from divergent to convergent systems challenging. We demonstrate through a set of case studies that it is imperative to represent model errors adequately and incorporate parameter uncertainties in ensemble data assimilation in convergent systems.

1 Introduction

Information on physical systems is often available in two forms: on the one hand from observations and on the other hand through mathematical models describing the systems dynamics. The combination of both can lead to an improved description of the system. This is the aim of data assimilation, typically with a focus on state estimation.

Data assimilation has broad applications throughout the geosciences and can be already seen as an independent discipline (Carrassi et al., 2018). It is typically used to estimate states, but also parameters: in weather forecasting (Houtekamer and Zhang, 2016; Ruiz et al., 2013), for atmospheric chemical transport (Carmichael et al., 2008; Zhang et al., 2012) also coupled to meteorology (Bocquet et al., 2015), in oceanography including biogeochemical processes (Stammer et al., 2016; Edwards et al., 2015), and in hydrology for flow, transport, and reaction in terrestrial surface and subsurface systems (Liu et al., 2012). Data assimilation is also increasingly applied in ecology with applications ranging from the spread of infectious diseases and wildfires, to population dynamics, and to the terrestrial carbon cycle (Niu et al., 2014; Luo et al., 2011).

In this study we distinguish geophysical systems between divergent and convergent systems, depending on the development of two initially nearby states (Fig. 1). In a divergent system, initially close states will inevitably drift apart, even if the system is described by a perfect model (Kalnay, 2003). This leads to an upper limit for the predictability in divergent systems (Lorenz,

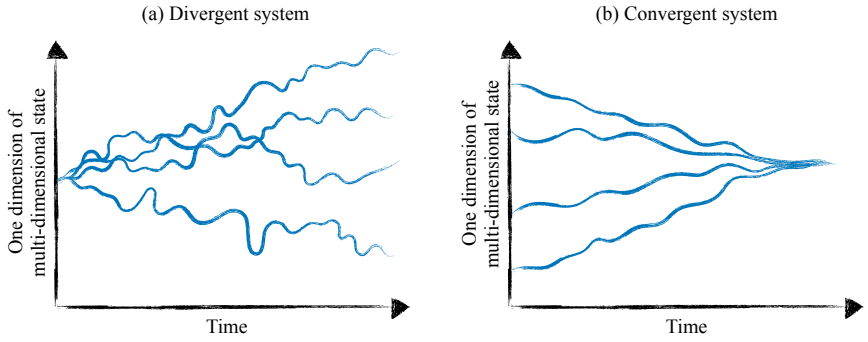

Figure 1. Dynamics of a generic divergent and convergent dynamic system with different initial states. Both panels show a single state dimension of a multi-dimensional system. In the divergent system, initially infinitesimally close states drift apart, while in the convergent system initially apart states converge.

1982). In a convergent system, nearby trajectories will coalesce. If the model to describe such a convergent system is perfect, this results in a high predictability (Lorenz, 1996). An error in the initial state will decay towards the truth after some transient phase. However, this is only true for perfect models, which is usually not the case for geophysical systems. This can lead to a bias with convergence to a wrong state.

Many recent advances of data assimilation methods have been developed in the context of weather forecasting (van Leeuwen et al., 2015). They are therefore designed to meet the challenges in the atmosphere – a predominantly divergent system. Due to the fundamental limit for long time predictions from uncertain initial conditions in divergent systems, data assimilation in operational weather forecasting primarily focuses on state estimation (Reichle, 2008; van Leeuwen et al., 2015).

While weather forecasting or oceanography are divergent systems, several geophysical system, such as soil hydrology or chemical transport are convergent systems. In ensemble data assimilation methods, if uncertainties are only represented through an ensemble of states, convergent systems lead to decreasing uncertainties over time, which favors filter degeneration. This difference to divergent systems, where the ensemble spread increases exponentially, makes the direct transfer of data assimilation methods from divergent systems to convergent systems challenging and often requires adaptations to prevent filter degeneracy. The application of data assimilation when coupling a divergent and a convergent model, for example in coupled chemistry meteorology models, may lead to potentially new difficulties (Bocquet et al., 2015).

The largest uncertainties in convergent systems typically do not reside in uncertain initial conditions but rather in boundary conditions, which include external forcings, the representation of sub-scale physics through parameterizations, and unrepresented physics in the model equations. These uncertainties should then be addressed integratedly (Liu and Gupta, 2007). Therefore, data assimilation methods have been used to not only estimate states but also parameters to reduce these uncertainties. The combined estimation of states and parameters is thought to be a solution of reducing the impact of model errors on parameter estimation (Liu et al., 2012). Estimating parameters in ensemble data assimilation methods through an augmented

state requires a forward model for the parameters as well. This model is typically assumed to be constant, which is neither divergent nor convergent. However, the filter will gradually reduce the uncertainty in the parameters, which is not increased through a divergent forward model and challenges similar to convergent systems can arise. This is sometimes alleviated by assigning a random walk as the forward model to the parameters, which then requires to determine an appropriate step size, however.

A challenge for sequential ensemble data assimilation in convergent systems it to maintain a sufficient ensemble spread. This would require an adequate representation of all uncertainties, including unrepresented physics in the model equations. In real world systems this is often difficult or impossible. Therefore, practical alternatives are necessary, which are often heuristic and may interfere with basic assumptions in the data assimilation methods, however. One possibility are inflation methods, which counteract the coalescing tendency. Unfortunately, there exist no universal method to accomplish this and a range of approaches are followed. One example is the increase of the ensemble spread of parameters to a threshold value, as soon as the parameter uncertainty drops below this value. This approach was introduced for a see-breeze model (Aksoy et al., 2006) and has been used in hydrology, as well (e.g. Shi et al., 2014; Rasmussen et al., 2015). A modification to this, also applied in hydrology, is to keep the parameter uncertainty entirely constant (Han et al., 2014; Zhang et al., 2017). This ensures a sufficient ensemble spread in the state itself, but can impact the accuracy of the estimation. A widely used adjustment to limit the reduction of an ensemble spread in hydrology is the use of a damping factor (Hendricks Franssen and Kinzelbach, 2008). In soil hydrology, a multiplicative inflation method was proposed, specifically adjusted to the needs of the system (Bauser et al., 2018). Similarly, Constantinescu et al. (2007) showed that an atmospheric chemical transport model required much stronger inflation than reported in the meteorological literature and showed better results for a model specific inflation, where the key parameters are perturbed to achieve an increased spread in the state. Consequently, a better understanding and control over errors has been recognized as a major challenge in chemical data assimilation as well (Zhang et al., 2012).

Although plenty of knowledge and experience is available in the different communities how to handle data assimilation methods in their specific model, we are not aware of a fundamental analysis of the difference between divergent and convergent models with respect to their utilization within ensemble data assimilation frameworks. We investigate and demonstrate the different challenges, that illustrate for example the different requirements for inflation methods, using the ensemble Kalman filter (Evensen, 1994; Burgers et al., 1998). The divergent case is illustrated using the Lorenz-96 model (Lorenz, 1996), while for the convergent case, a soil hydrological system described by Richards' equation is used. Naturally, these are highly simplified models compared to real-world applications. Still, they demonstrate key aspects that also have to be addressed in more complicated situations. The specific adjustments applied there depend on the particular model. Our focus here is on the fundamentals, not on the wide range of specifics.

2 Data assimilation method

For this study we chose the ensemble Kalman filter (EnKF), a sequential data assimilation method, based on Bayes' theorem and the assumption of unbiased Gaussian error distributions. It was introduced as an extension of the Kalman filter (Kalman,

1960) for nonlinear models (Evensen, 1994; Burgers et al., 1998) and approximates the Gaussian distributions by an ensemble of states $\boldsymbol{\psi}_i$, where the subscript i denotes the ensemble member.

To sequentially assimilate new observations, the EnKF alternates between a forecast (subscript 'f') and an analysis step (subscript 'a'). In the forecast, the ensemble is propagated using the nonlinear model $f(\cdot)$

$$\boldsymbol{\psi}_{\mathrm{f},i}^k = f(\boldsymbol{\psi}_{\mathrm{f},i}^{k-1}) + \boldsymbol{\beta}^k \,, \tag{1}$$

where $\boldsymbol{\beta}^k$ is a stochastic model error and k is a discrete time.

The analysis state $\boldsymbol{\psi}_\mathrm{a}$ is calculated by applying the Kalman gain

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

to every ensemble member. The Kalman gain weights the forecast error covariance $\mathbf{P}$ with the observation error covariance $\mathbf{R}$. The observation operator $\mathbf{H}$ maps the state from state space to observation space. The forecast error covariance $\mathbf{P}$ is calculated using the forecast ensemble

$$\mathbf{P}_\mathrm{f}^k = \overline{\left(\boldsymbol{\psi}_\mathrm{f}^k - \overline{\boldsymbol{\psi}_\mathrm{f}^k}\right)\left(\boldsymbol{\psi}_\mathrm{f}^k - \overline{\boldsymbol{\psi}_\mathrm{f}^k}\right)^\mathsf{T}} \,. \tag{3}$$

It is necessary to add a realization of the observation error to the observation for each ensemble member (Burgers et al., 1998). The resulting analysis state is

$$\boldsymbol{\psi}_{\mathrm{a},i}^k = \boldsymbol{\psi}_{\mathrm{f},i}^k + \mathbf{K}^k \left(\boldsymbol{d}^k - \mathbf{H}\boldsymbol{\psi}_{\mathrm{f},i}^k + \boldsymbol{\epsilon}_i^k\right) \,, \quad \epsilon_i^k \propto \mathcal{N}(\mathbf{0}, \mathbf{R}^k) \,. \tag{4}$$

By combining the information from measurement and model, the uncertainty in the analysis ensemble is decreased. For sequential data assimilation, the process of forecast and analysis is iterated for every new observation.

The joint estimation of states and parameters can be realized through state augmentation. The original state $\boldsymbol{\psi}$ is extended with the parameters $\boldsymbol{p}$ to an augmented state

$$\boldsymbol{u} = \begin{pmatrix} \boldsymbol{\psi} \\ \boldsymbol{p} \end{pmatrix} . \tag{5}$$

The model equation (Eq. 1) changes to

$$\boldsymbol{u}^k = \begin{pmatrix} \boldsymbol{\psi}^k \\ \boldsymbol{p}^k \end{pmatrix} = \begin{pmatrix} f_{\boldsymbol{\psi}}\left(\boldsymbol{\psi}^{k-1}, \boldsymbol{p}^{k-1}\right) + \boldsymbol{\beta}_{\boldsymbol{\psi}}^k \\ f_{\boldsymbol{p}}\left(\boldsymbol{p}^{k-1}\right) + \boldsymbol{\beta}_{\boldsymbol{p}} \end{pmatrix} , \tag{6}$$

with the models for the state $f_{\boldsymbol{\psi}}(\cdot)$ and for the parameters $f_{\boldsymbol{p}}(\cdot)$, as well as the corresponding model errors $\boldsymbol{\beta}_{\boldsymbol{\psi}}$ and $\boldsymbol{\beta}_{\boldsymbol{p}}$. The model for the state is typical nonlinear, while the parameters are often assumed constant $f_{\boldsymbol{p}}(\boldsymbol{p}^{k-1}) = \boldsymbol{p}^{k-1}$.

In this study, we set both stochastic model errors in Eq. (6) to zero. The EnKF is used without any extensions and an ensemble size of $N = 100$ was chosen for all cases.

3 Divergent system

This section demonstrates the data assimilation behavior for a divergent system on the example of the 40-dimensional Lorenz-96 model, which has been widely used to test data assimilation methods in atmospheric sciences (e.g., Li et al., 2009). We first introduce the model before we look into four characteristic different cases.

3.1 Lorenz-96

The Lorenz-96 model (Lorenz, 1996) is an artificial model and cannot be derived from any dynamic equation (Lorenz, 2005). It can be interpreted as an unspecified scalar quantity x in a one dimensional atmosphere on a latitude circle and was defined in a study on predictability (Lorenz, 1996).

The governing equations are a set of coupled ordinary differential equations:

$$\frac{dx_i}{dt} = (x_{i+1} - x_{i-2})x_{i-1} - x_i + F \quad , \quad i \in [1, 2, \ldots, J] \tag{7}$$

with constant forcing F, periodic boundaries ($x_{J+1} = x_1$) and dimension J. The dimension is often chosen as $J = 40$, which we also do in this study. Even though the system is not derived from physical principles, it shares certain properties of large atmospheric models (Lorenz and Emanuel, 1998). The quadratic terms represent advection and conserve the total energy, while the linear term decreases the total energy comparable to dissipation. The constant F represents external forcing and prevents the system's total energy from decaying to zero. The value is often chosen as $F = 8$ (Lorenz and Emanuel, 1998).

The Lyapunov exponent quantifies how fast two initially infinitesimal close trajectories will separate. Analyzing the leading Lyapunov exponent for $F = 8$ shows a doubling time of $\tau_d = 0.42$ units for the distance between two initially infinitesimally close neighboring states (Lorenz and Emanuel, 1998). Increasing the forcing leads to a more divergent system, for instance $\tau_d = 0.3$ for $F = 10$ (Lorenz, 1996). With τ_d decreasing, so does the predictability of the system.

3.2 Characteristic cases

The behavior of the EnKF on a divergent system is investigated through four different cases (DC1-4), which are purposely designed simple to illustrate the behavior concisely. For all cases, the model is solved using a fourth order Runge-Kutta method with a time step of $\Delta t = 0.01$.

The initial condition for the synthetic truth for all cases is generated by running the model until time 2000 from an initial state $x_i = 4.0 \; \forall i \in [1, 2, \ldots, 39]$ and $x_{40} = 4.001$, with the typical value $F = 8$ for the forcing parameter. The final state of this run is used as the synthetic true initial state for all cases. This ensures that the state is on the attractor without the initial transient phase. The initial ensemble for the data assimilation runs is generated by perturbing the true initial state with a Gaussian distribution $\mathcal{N}(0, 1)$.

Synthetic observations are generated in all 40 dimensions by a forward run until time 4, using the true value and perturbing it with a Gaussian distribution with a zero mean and a standard deviation of $\sigma_{\text{obs}} = 1.0$. For cases DC1 and DC2 observations are generated at 8 different times with an observation interval of $\Delta t_{\text{Obs}} = 0.5$. This observation interval is chosen rather large

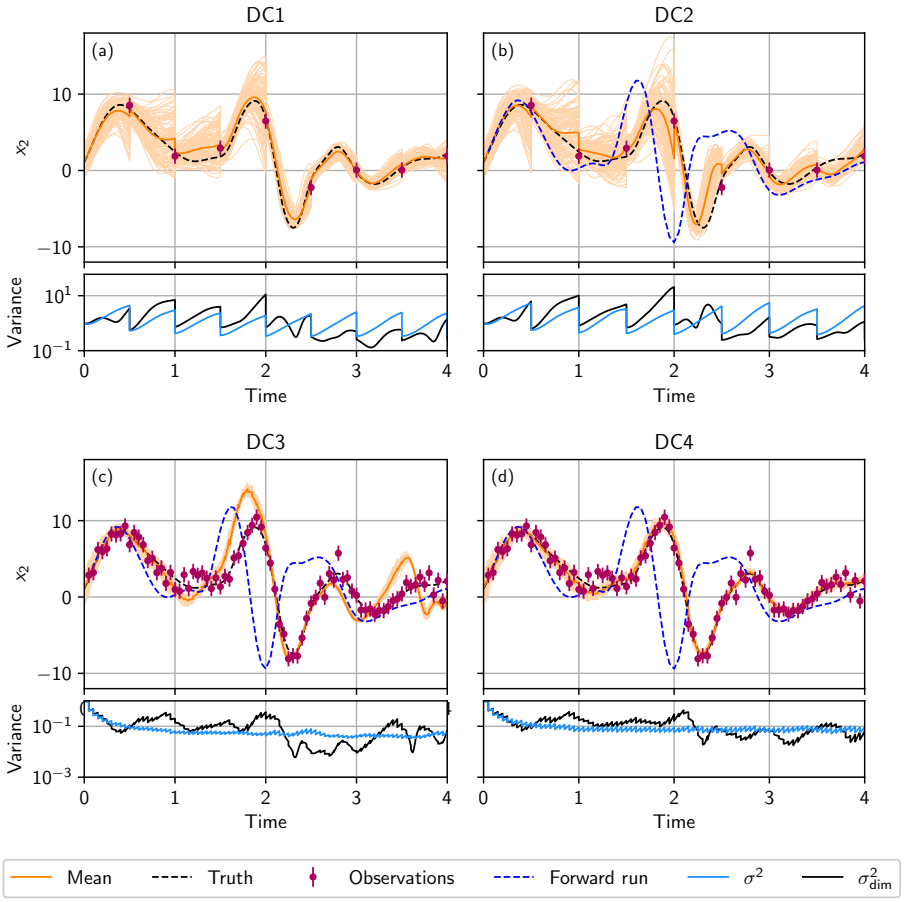

Figure 2. State estimation in a divergent system. (a): Divergent case 1 (DC1), the ensemble is propagated with the same parameter as the truth ($F = 8$). (b): Divergent case 2 (DC2), the ensemble is propagated with $F = 10$ instead. (c): Divergent case 3 (DC3), the ensemble is propagated with $F = 10$ and the observation interval is reduced to $\Delta t_{\text{Obs}} = 0.05$. (d): Divergent case 4 (DC4), the parameter error is represented by the ensemble using $\mathcal{N}(10, 2^2)$ for the reduced observation interval of $\Delta t_{\text{Obs}} = 0.05$. Top panels : The ensemble mean (orange line) and the ensemble (light orange lines) in the data assimilation run for state dimension 2 x_2 together with the observations (purple), generated from the truth (black dashed line). (b-d) additionally show a single forward run (blue dashed line) using a wrong parameter $F = 10$ starting from the true initial condition. Bottom panels: Mean variance σ^2 of the ensemble over all dimensions (light blue line) and variance σ^2_{dim} of state dimension 2 (black line).

135 to ensure a large divergence of the system. For cases DC3 and DC4 observations are generated at 80 different times with an observation interval of $\Delta t_{\text{Obs}} = 0.05$. This interval is often used in other studies (e.g. Nakano et al., 2007; van Leeuwen, 2010; Poterjoy, 2016).

3.2.1 Divergent case 1 (DC1) – state estimation, true parameter

In this case, the ensemble is propagated with the same parameter ($F = 8$) as the synthetic truth.Uncertainty only stems from the uncertainty in the initial condition. Figure 2a shows the time development of one state dimension of the Lorenz-96 model (top panel) and the ensemble variance of this state dimension σ^2_{dim} together with the mean variance over all dimensions σ^2 (bottom panel). Due to the divergent nature of the model, the ensemble spread increases between observations and the ensemble has a sufficient spread such that the EnKF is able to correct the states to follow the truth.

The mean ensemble spread σ^2 increases exponentially between two observations. At each observation, the EnKF updates the ensemble and the variance decreases correspondingly.

The behavior of σ^2_{dim} differs from the behavior of σ^2. During the forecast, σ^2_{dim} sometimes increases, decreases or stays approximately constant. This occurs because the Lorenz-96 model is bounded and has, therefore, convergent and divergent directions.

3.2.2 Divergent case 2 (DC2) – state estimation, wrong parameter

In this case, in order to investigate the impact of an unrepresented model error, the ensemble is propagated with a wrong parameter of $F = 10$ instead of $F = 8$ (Fig. 2b), which was used to generate the observations. Therefore, the ensemble is propagated with a model that is more divergent than the synthetic truth.

Compared to DC1 (Fig. 2a, bottom panel), σ^2 and σ^2_{dim} increase faster and the ensemble spread reaches higher values, which shows the increased divergence of the system. Propagating the ensemble with this different model leads to a larger deviation of the ensemble mean from the truth (see Fig. 2b, top panel) than in DC1. However, the divergent nature of the model ensures a sufficient ensemble spread such that the state can be corrected towards the truth and the wrong parameter is compensated. This leads to a worsened but still good estimation of the truth without filter divergence.

3.2.3 Divergent case 3 (DC3) – state estimation, small observation interval, wrong parameter

The case is similar to DC2 except that the observation interval is reduced to $\Delta t_{\text{Obs}} = 0.05$. The high frequency of analysis steps, where the Kalman update reduces the variance in the ensemble, prevents the Lorenz-96 model to develop enough divergence to increase the ensemble spread sufficiently in the short time intervals in between (see Fig. 2c, bottom panel). Although the ensemble is propagated with a more divergent model ($F = 10$), the divergence is not sufficient to encompass the model error due to the wrong parameter and the filter can degenerate (see Fig. 2c, top panel). This is in contrast to DC2, where the Kalman filter can successfully estimate the state. However, the comparison with the forecast of the true initial state using the wrong parameter shows that the EnKF is able to improve the state significantly.

3.2.4 Divergent case 4 (DC4) – state estimation, small observation interval, represented error

In divergent case 4 (DC4), the observation interval is, as in DC3, reduced to $\Delta t_{\text{Obs}} = 0.05$. The parameter error is represented by the ensemble by assigning each ensemble member a different parameter F. The forcing parameter F is drawn from a Gaussian distribution $\mathcal{N}(10, 2^2)$, such that the true value lies within one standard deviation .

Representing the error increases σ^2 only slightly (Fig. 2d, bottom panel) compared to DC3, but the minimal variance of σ^2_{dim} has higher values. This increase is sufficient to help the filter such that it does not degenerate (Fig. 2d, top panel). Representing the parameter error in the ensemble in the case of frequent observations can prevent filter degeneracy.

4 Convergent system

This section demonstrates the data assimilation behavior for a convergent system on the example of soil water flow. We again first introduce the model before we look into four characteristic cases.

4.1 Soil water flow

Water flow in an unsaturated porous medium can be described by the Richards equation:

$$\partial_t \theta - \nabla \cdot [K(\theta)\,[\nabla h_m(\theta) - 1]] = 0\,, \tag{8}$$

where θ $(-)$ is the volumetric water content, K $(\mathrm{L}\,\mathrm{T}^{-1})$ is the isotropic hydraulic conductivity, and h_m (L) is the matric head.

To close Eq. 8 soil hydraulic material properties are necessary, which specify the dependency of the matric head and the hydraulic conductivity on the water content. We use the Mualem-van Genuchten parametrisation (Mualem, 1976; Van Genuchten, 1980) in its simplified form:

$$K(\Theta) = K_w\,\Theta^{\tau}\left[1 - \left[1 - \Theta^{n/[n-1]}\right]^{1-1/n}\right]^2\,, \tag{9}$$

$$h_m(\Theta) = \frac{1}{\alpha}\left[\Theta^{-n/[n-1]} - 1\right]^{1-1/n}\,, \tag{10}$$

with the saturation Θ $(-)$

$$\Theta := \frac{\theta - \theta_r}{\theta_s - \theta_r}\,. \tag{11}$$

The parameter θ_s $(-)$ is the saturated water content and θ_r $(-)$ is the residual water content. The matric head h_m is scaled with the parameter α (L^{-1}) that can be related to the inverse air entry value. The parameter K_w $(\mathrm{L}\,\mathrm{T}^{-1})$ is the saturated hydraulic conductivity, τ $(-)$ a tortuosity factor and n $(-)$ is a shape parameter. Equation (9) and Eq. (10) describe the sub-scale physics with six parameters for a homogeneous soil.

4.2 Characteristic cases

The behavior of the EnKF on a convergent system is investigated through four different cases (CC1-4). For the case studies, a one-dimensional homogeneous soil is used with an extent of 1 m. The Richards equation is solved using MuPhi (Ippisch et al., 2006) with a spatial resolution of 1 cm, which results in a 100-dimensional water content state.

For the true trajectories and the observations, parameters for a loamy sand by Carsel and Parrish (1988) are used: $\theta_s = 0.41$, $\theta_r = 0.057$, $\tau = 0.5$, $n = 2.28$, $\alpha = -12.4\,\mathrm{m}^{-1}$, and $K_w = 4.00 \cdot 10^{-5}\,\mathrm{m\,s}^{-1}$. For the lower boundary, a Dirichlet condition with zero potential (groundwater table) is set and for the upper boundary a constant infiltration over the whole observation time with a flux of $5 \cdot 10^{-7}\,\mathrm{m\,s}^{-1}$ is used.

Initially, the system is in hydraulic equilibrium. The infiltration boundary condition leads to a downward propagating infiltration front increasing the water content. Four time domain reflectometry (TDR)-like water content observations are generated equidistantly at depths of $(0.2, 0.4, 0.6, 0.8)$ m. The observation error is chosen to be $\sigma_{\mathrm{Obs}} = 0.007$ (e.g., Jaumann and Roth, 2017). Observations are taken hourly for a duration of 30 h.

To generate the initial ensemble, the ensemble mean of the water content state is perturbed by a correlated multivariate Gaussian distribution using a Cholesky decomposition to create an ensemble that corresponds to a predefined covariance matrix (e.g., Berg et al., 2019). The main diagonal of this covariance matrix of the ensemble is 0.003^2. The off-diagonal entries are determined by multiplying the variance on the main diagonal with the the fifth-order piecewise rational function by Gaspari and Cohn (1999) using a length scale of $c = 10$ cm. This ensures a spatially correlated initial state, which increases the diversity of the ensemble. If instead uncorrelated Gaussian random numbers with a zero mean were used, the dissipative component of the system would lead to a fast vanishing of the perturbation in space for each individual ensemble member.

4.2.1 Convergent case 1 (CC1) – no estimation

In this case no data assimilation is used and the ensemble is propagated with the true model. The initial conditions for the ensemble members are based on the linearly interpolated observations at time zero. This approximated state is used as the ensemble mean for the EnKF. This state is then perturbed by a correlated multivariate Gaussian distribution such that the spread of the initial ensemble is sufficient to represent the uncertainty of the water content in most parts.

The temporal development of the water content at 20 cm depth, the position of the uppermost observation, is shown in Fig. 3a (top panel). Due to the convergent behavior of the model in combination with the true representation, the initially broad ensemble converges to the truth, even though the initial condition was not represented accurately.

The ensemble variance at this depth σ^2_{dim} increases when the infiltration front reaches 20 cm. Because of the nonlinear conductivity function (Eq. 9), the different initial water contents lead to a different arrival time of the infiltration front at the first observation position. This leads to an increase in the ensemble spread.

After the increase of the water content, the ensemble collapses fast since the hydraulic conductivity increases with the water content, which leads to a fast convergence of the different ensemble members to the truth due the convergent nature of the

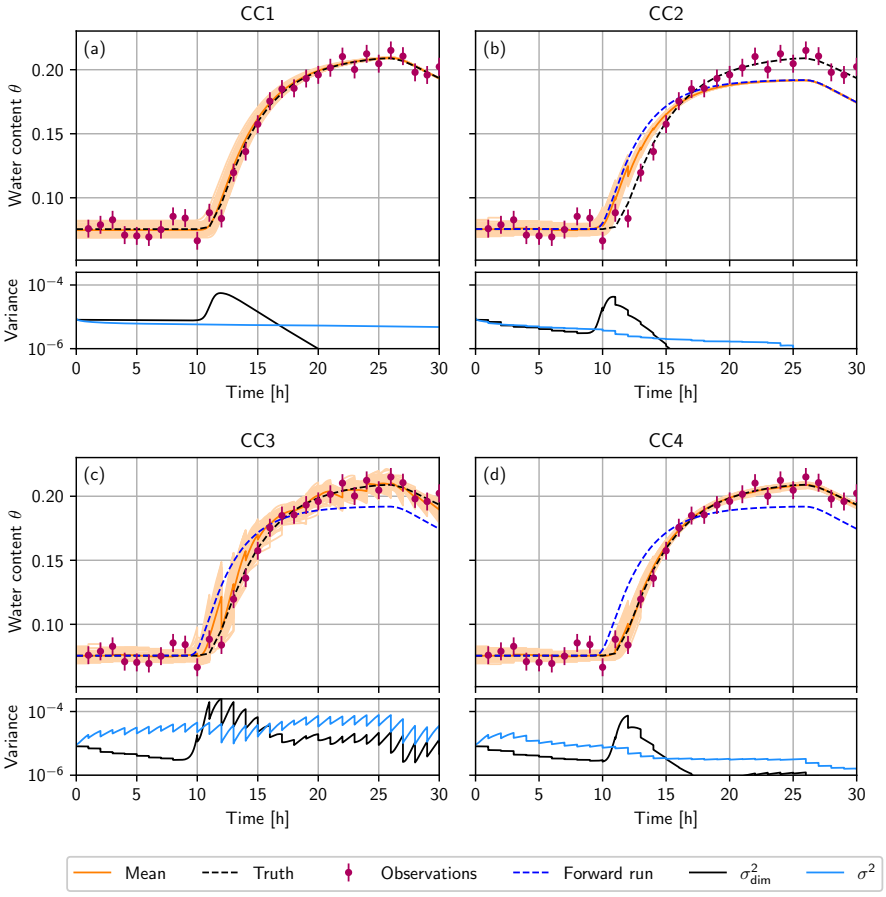

Figure 3. (a): Convergent case 1 (CC1): forward run without data assimilation, using an interpolated initial condition. In the other three cases the truth as used to generate the initial ensemble. (b): Convergent case 2 (CC2): the ensemble is propagated with $n = 2.68$ instead of $n_{\text{true}} = 2.28$ and the state is estimated. (c): Convergent case 3 (CC3): the parameter error is represented by the ensemble using $\mathcal{N}(2.68, 0.4^2)$ and the state is estimated. (d): Convergent case 4 (CC4): simultaneous state and parameter estimation. Top panels: The ensemble mean (orange) and the ensemble (light orange lines) during the forward run at the depth of the uppermost observation (20 cm). The truth, which is used to generate the observations (purple), is shown as a black dashed line. (b-d) additionally show a single forward run (blue dashed line) using a wrong parameter $n = 2.68$ starting from the true initial condition. Bottom panels: Mean variance σ^2 of the ensemble over all dimensions (light blue line) and variance σ^2_{dim} at the depth of 20 cm (black line).

model. The variance over all dimensions σ^2 decreases slowly and approaches zero over time. If the system is started with a higher water content instead of equilibrium, this collapse will occur faster.

225 This case shows that a perfectly convergent system is predictable for all times. This is in contrast to divergent systems, that only have a finite limit of predictability even if the model including boundary conditions is perfectly known. After a transient phase, the states converge to the truth (Kalnay, 2003). A perfect model is not what we encounter in reality, however.

4.2.2 Convergent case 2 (CC2) – state estimation, wrong parameter

In this case, the state is estimated with the EnKF but the ensemble is propagated with a wrong parameter. Instead of $n_{\text{true}} = 2.28$, $n = 2.68$ is chosen. The mean for the initial state is chosen as the true initial water content.

In Fig. 3b (top panel) the temporal development of the water content at the depth of the uppermost observation (20 cm) is shown. A larger n results in an earlier arrival of the infiltration front at the depth of 20 cm for the ensemble than for the truth. The EnKF tries to correct the ensemble but fails because its variance is too small and cannot represent the truth. Due to the convergent system, σ^2 decreases constantly while σ^2_{dim} decreases fast to zero after the infiltration front reaches the depth of 20 cm (see Fig. 3b, bottom panel). This convergence leads to a false trust in the model and the filter degenerates. Compared to a forward run without data assimilation, the EnKF can only improve the state estimation for a short time when the water content rises due to the infiltration front. Soon after, the ensemble coincides with the free forward run and the estimated state shows no advantage any more.

This case illustrates that a wrong parameter in a convergent system can lead to filter degeneration. This is in direct contrast to DC2 (Fig. 2b), where the filter is still able to estimate the state. The behavior also differs from DC3 (Fig. 2c), where the observation interval in the Lorenz-96 model is too short such that it cannot develop its full divergent behavior. There the filter can also degenerate for a wrong parameter, but the data assimilation is still able to improve the estimation compared to a free forward run, since the ensemble never collapses entirely.

4.2.3 Convergent case 3 (CC3) – state estimation, represented error

In this case, the parameter error is represented with the ensemble but the parameter is not estimated. Each ensemble member has a different parameter n. The parameters are Gaussian distributed with $\mathcal{N}(2.68, 0.4^2)$ such that the truth lies within one standard deviation. Since the model error is known in this synthetic case we can create an ensemble that represents the model error adequately. Note, that this is more difficult or even impossible in a real-world system. The mean for the initial state is chosen as the true initial water content.

Figure 3c (top panel) shows the temporal development of the water content at a depth of 20 cm. The infiltration front reaches this depth at different times due to the different parameter n for each ensemble member. This increases the variance in the ensemble, both at this depth and overall (see Fig. 3c, bottom panel). The variance increases rapidly between the observations, similar to the divergent cases. This way, the ensemble spread stays large enough such that the EnKF can correct the states. The ensemble can follow and represent the truth. This behavior can also be observed for the divergent case DC4 with a short observation interval (Fig. 2d).

Representing the model error adds a divergent component to the ensemble for a convergent model. This allows the EnKF to correct the state and follow the truth. However, the predictability of the system decreases since each ensemble member converges to a different fixed point apart from the truth. To increase the predictability, parameter estimation is necessary.

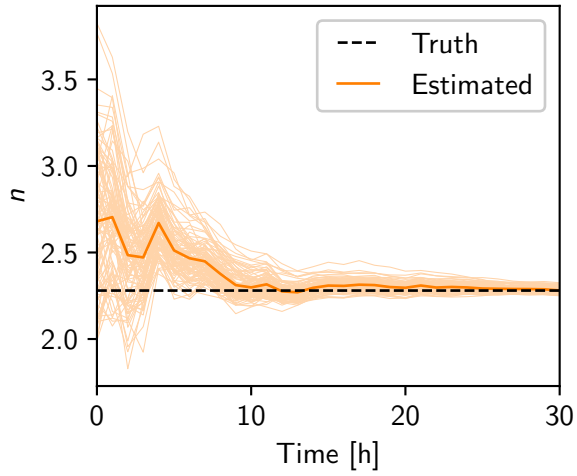

Figure 4. Estimation of parameter n in convergent case 4 (CC4). The ensemble mean is shown in orange and the ensemble in light orange. The truth is a black dashed line.

4.2.4 Convergent case 4 (CC4) – state and parameter estimation

In this case, the error in the parameter n is not only represented but also estimated using the state augmentation method. The initial parameter set is, as in CC3, Gaussian distributed with $\mathcal{N}(2.68, 0.4^2)$ such that the truth is located within one standard deviation. Again, the mean for the initial state is chosen as the true initial water content.

The estimation of n is shown in Fig. 4. The ensemble converges rapidly to the truth because only one parameter is estimated, so every deviation from the truth is mainly caused by this parameter.

The mean variance σ^2 increases initially (Fig. 3d, bottom panel), because in the beginning the parameter has not been sufficiently improved such that the ensemble members still have different n. This leads to a divergent ensemble in state space during the infiltration similar to CC3. While the parameter is estimated, the variance of the ensemble decreases fast and the convergent property of the system becomes dominant.

The temporal development of the water content at a depth of 20 cm is shown in Fig. 3d (top panel). In contrast to CC3, the corrections of the EnKF to the state are much smaller. The mean of the parameter n comes close to the true value and the uncertainty of n decreases. This causes the forward propagation to come close to the true model as well. The propagation with an almost correct model supports the state estimation due to the convergent nature of the system which forces the state to the true value.

5 Discussion

For the divergent Lorenz-96 system, the EnKF is able to estimate the state for the true model as well as for the case with a wrong parameter. In a divergent system, the volume of the prior in state space increases during forward propagation (Evensen, 1994). For the EnKF, this is directly connected to the ensemble spread, which increases rapidly between the observations. This prevents a collapse of the ensemble even in the presence of an unrepresented parameter error. However, if the observation interval is too small relative to the characteristic time for divergence, the EnKF leads to a decrease in the ensemble spread such that the filter degenerates in the case of an unrepresented parameter error. Nevertheless, the divergent behavior of the model prevents a complete collapse of the ensemble, so that the filter is still able to improve the state in a limited way.

In the convergent soil hydrological system, the volume of the prior distribution decreases during forward propagation such that the prior becomes more certain even without an observation and data assimilation. For a perfect model, the predictability and state estimation in a convergent model are trivial. The initial ensemble will converge to the truth after some time, even with a rough initial approximation. In this case, data assimilation is not necessary.

In the case of model errors, in our study realized through a wrong parameter, the situation is different. The ensemble converges to a wrong state, the filter degenerates and data assimilation fails. Increasing the ensemble size can only improve the performance marginally since all ensemble members converge to the same fixed point.

Representing the parameter error by assigning each ensemble member a different parameter, increases the divergence of the system and the filter is able to estimate the state again. Between the observations, the ensemble spread increases rapidly because the ensemble members diverge to different fixed points apart from the truth. This results in a finite predictability. By representing the parameter error, the Richards equation gains a divergent part similar to the Lorenz-96 model. In the case of a convergent system, it is necessary to represent the parameter error, otherwise the ensemble collapses.

To increase the predictability of the system again, it is necessary to not only represent but also to reduce the parameter uncertainty. In synthetic cases without model structural errors, the convergent property of the system supports the state estimation and the predictability increases, if the parameter estimation is successful. This shows the importance of parameter estimation for convergent systems. For divergent systems parameter estimation also increases the predictability but only up to a point, because predictability is limited by the system's dynamics.

For the application of data assimilation to real data, model errors typically cannot be attributed to unknown parameters or uncertainties in boundary conditions alone, but also stem from model structural errors like a simplified representation of sub-scale physics or unrepresented processes in the dynamics. Uncertainties in parameters and boundary conditions can be represented in an ensemble, but representing model structural errors is challenging or can be impossible. For example, in hydrology, the model errors are typically ill-known (Li and Ren, 2011) and can vary both in space and time, which then can lead to filter degeneracy and biased parameter estimation (Berg et al., 2019). While divergent models can alleviate the effect of an unrepresented error within bounds, in a convergent system it is necessary to represent all relevant model errors sufficiently to prevent filter degeneracy and enable an optimal state and parameter estimation. Since this is challenging or even impossible, heuristic ways to address filter degeneracy are necessary.

A practical alternative to increase the ensemble spread and avoid filter degeneracy are inflation methods. For example, keeping a constant ensemble spread for the parameters can provide a sufficient spread in state space (Zhang et al., 2017). The advantage is that this approach is a model-specific inflation and avoids overamplification of spurious correlations (Constantinescu et al., 2007). The disadvantage is, that the EnKF is prevented from reducing the prediction uncertainty. This behavior is also shown in CC3. In this case, the ensemble spread in the parameter space adds a divergent component to the ensemble that results in an increased ensemble spread in state space and prevents filter degeneracy.

Multiplicative inflation increases the ensemble spread through an inflation factor (Anderson and Anderson, 1999). In a divergent system a small multiplicative inflation is sufficient to increase the existing ensemble spread. In a convergent system this requires a larger inflation and can lead to overinflation of spurious correlations (Constantinescu et al., 2007). To avoid this and to cope with spatial and temporal varying model errors, the use of more sophisticated adaptive inflation methods (e.g. Bauser et al., 2018; Gharamti, 2018) may be necessary in convergent systems.

Heuristic inflation methods cannot fully replace the representation of model errors and hence must be used judiciously. They can lead to overinflation of spurious correlations and can lead to biases in the estimation of parameters. If unrepresented model errors can be identified and are limited in space or time, these biases can be prevented by using a closed-eye period for the extent of the unrepresented model errors. In the closed-eye period, parameters are kept constant and only the state is estimated, which can require inflation methods for a successful state estimation. This enables an improved parameter estimation without compensating the unrepresented model dynamics though biased parameters outside of the closed-eye period, when and where uncertainties can be represented. The use of the closed-eye period in combination with the representation of relevant uncertainties has already been demonstrated in hydrology (Bauser et al., 2016).

6 Conclusions

We demonstrated the difference and challenges of ensemble data assimilation for divergent and convergent systems on the example of the EnKF applied to the divergent Lorenz-96 model and a convergent soil water movement model based on the Richards equation.

Sequential ensemble data assimilation methods require a sufficient divergent part in the ensemble to maintain an adequate ensemble spread and prevent filter degeneration. In divergent systems this is inherent to the system, provided that observation intervals and divergence times match. In convergent systems relevant model errors must be represented to increase the ensemble spread and to thereby add a divergent part to the ensemble to avoid filter degeneracy. If errors stem from unknown parameters, estimating the parameters improves state estimation. However, this will reduce the ensemble spread again and require the remaining relevant model errors to be represented. Since this can be challenging, increasing the model errors artificially, by limiting the reduction of parameter uncertainty or through inflation methods, can be required in convergent systems.

This paper highlights the challenges when transferring sequential ensemble data assimilation methods from divergent systems to convergent systems, which must be considered when applying data assimilation.

Author contributions. HHB and DB designed and implemented the presented study. DB performed and analyzed the simulations. All authors participated in continuous discussions. HHB and DB prepared the manuscript with contributions from KR. The presented results are based on the PhD thesis by DB.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This research was funded by Deutsche Forschungsgemeinschaft (DFG) through project RO 1080/12-1. Daniel Berg was supported in part by the Heidelberg Graduate School of Mathematical and Computational Methods for the Sciences (HGS MathComp), funded by DFG grant GSC 220 in the German Universities Excellence Initiative. Hannes H. Bauser was supported in part by the Deutsche Forschungsgemeinschaft (DFG) through Project BA 6635/1-1.

References

Aksoy, A., Zhang, F., and Nielsen-Gammon, J. W.: Ensemble-based simultaneous state and parameter estimation in a two-dimensional sea-breeze model, Monthly Weather Review, 134, 2951–2970, https://doi.org/10.1175/MWR3224.1, 2006.

Anderson, J. L. and Anderson, S. L.: A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts, Monthly Weather Review, 127, 2741–2758, https://doi.org/10.1175/1520-0493(1999)127<2741:AMCIOT>2.0.CO;2, 1999.

Bauser, H. H., Jaumann, S., Berg, D., and Roth, K.: EnKF with closed-eye period – towards a consistent aggregation of information in soil hydrology, Hydrology and Earth System Sciences, 20, 4999–5014, https://doi.org/10.5194/hess-20-4999-2016, 2016.

Bauser, H. H., Berg, D., Klein, O., and Roth, K.: Inflation method for ensemble Kalman filter in soil hydrology, Hydrology and Earth System Sciences, 22, 4921–4934, https://doi.org/10.5194/hess-22-4921-2018, 2018.

Berg, D., Bauser, H. H., and Roth, K.: Covariance resampling for particle filter – state and parameter estimation for soil hydrology, Hydrology and Earth System Sciences, 23, 1163–1178, https://doi.org/10.5194/hess-23-1163-2019, 2019.

Bocquet, M., Elbern, H., Eskes, H., Hirtl, M., Žabkar, R., Carmichael, G. R., Flemming, J., Inness, A., Pagowski, M., Pérez Camaño, J. L., Saide, P. E., San Jose, R., Sofiev, M., Vira, J., Baklanov, A., Carnevale, C., Grell, G., and Seigneur, C.: Data assimilation in atmospheric chemistry models: current status and future prospects for coupled chemistry meteorology models, Atmospheric Chemistry and Physics, 15, 5325–5358, https://doi.org/10.5194/acp-15-5325-2015, 2015.

Burgers, G., van Leeuwen, P. J., and Evensen, G.: Analysis scheme in the ensemble Kalman filter, Monthly Weather Review, 126, 1719–1724, https://doi.org/10.1175/1520-0493(1998)126<1719:ASITEK>2.0.CO;2, 1998.

Carmichael, G. R., Sandu, A., Chai, T., Daescu, D. N., Constantinescu, E. M., and Tang, Y.: Predicting air quality: Improvements through advanced methods to integrate models and measurements, Journal of Computational Physics, 227, 3540 – 3571, https://doi.org/https://doi.org/10.1016/j.jcp.2007.02.024, predicting weather, climate and extreme events, 2008.

Carrassi, A., Bocquet, M., Bertino, L., and Evensen, G.: Data assimilation in the geosciences: An overview of methods, issues, and perspectives, Wiley Interdisciplinary Reviews: Climate Change, 9, e535, https://doi.org/10.1002/wcc.535, 2018.

Carsel, R. F. and Parrish, R. S.: Developing joint probability distributions of soil water retention characteristics, Water Resources Research, 24, 755–769, https://doi.org/10.1029/WR024i005p00755, 1988.

Constantinescu, E. M., Sandu, A., Chai, T., and Carmichael, G. R.: Ensemble-based chemical data assimilation. I: General approach, Quarterly Journal of the Royal Meteorological Society, 133, 1229–1243, https://doi.org/10.1002/qj.76, 2007.

Edwards, C. A., Moore, A. M., Hoteit, I., and Cornuelle, B. D.: Regional ocean data assimilation, Annual Review of Marine Science, 7, 21–42, https://doi.org/10.1146/annurev-marine-010814-015821, 2015.

Evensen, G.: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, Journal of Geophysical Research: Oceans, 99, 10 143–10 162, https://doi.org/10.1029/94JC00572, 1994.

Gaspari, G. and Cohn, S. E.: Construction of correlation functions in two and three dimensions, Quarterly Journal of the Royal Meteorological Society, 125, 723–757, https://doi.org/10.1002/qj.49712555417, 1999.

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

Han, X., Hendricks Franssen, H.-J., Montzka, C., and Vereecken, H.: Soil moisture and soil properties estimation in the Community Land Model with synthetic brightness temperature observations, Water Resources Research, 50, 6081–6105, https://doi.org/10.1002/2013WR014586, 2014.

Hendricks Franssen, H. J. and Kinzelbach, W.: Real-time groundwater flow modeling with the ensemble Kalman filter: Joint estimation of states and parameters and the filter inbreeding problem, Water Resources Research, 44, https://doi.org/10.1029/2007WR006505, 2008.

Houtekamer, P. L. and Zhang, F.: Review of the ensemble Kalman filter for atmospheric data assimilation, Monthly Weather Review, 144, 4489–4532, https://doi.org/10.1175/MWR-D-15-0440.1, 2016.

Ippisch, O., Vogel, H.-J., and Bastian, P.: Validity limits for the van Genuchten–Mualem model and implications for parameter estimation and numerical simulation, Advances in Water Resources, 29, 1780–1789, https://doi.org/10.1016/j.advwatres.2005.12.011, 2006.

Jaumann, S. and Roth, K.: Effect of unrepresented model errors on estimated soil hydraulic material properties, Hydrology and Earth System Sciences, 21, 4301–4322, https://doi.org/10.5194/hess-21-4301-2017, 2017.

Kalman, R. E.: A new approach to linear filtering and prediction problems, Journal of basic Engineering, 82, 35–45, https://doi.org/10.1115/1.3662552, 1960.

Kalnay, E.: Atmospheric modeling, data assimilation and predictability, Cambridge university press, 2003.

Li, C. and Ren, L.: Estimation of unsaturated soil hydraulic parameters using the ensemble Kalman filter, Vadose Zone Journal, 10, 1205, https://doi.org/10.2136/vzj2010.0159, 2011.

Li, H., Kalnay, E., and Miyoshi, T.: Simultaneous estimation of covariance inflation and observation errors within an ensemble Kalman filter, Quarterly Journal of the Royal Meteorological Society, 135, 523–533, https://doi.org/10.1002/qj.371, 2009.

Liu, Y. and Gupta, H. V.: Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework, Water Resources Research, 43, https://doi.org/10.1029/2006WR005756, w07401, 2007.

Liu, Y., Weerts, A. H., Clark, M., Hendricks Franssen, H.-J., Kumar, S., Moradkhani, H., Seo, D.-J., Schwanenberg, D., Smith, P., van Dijk, A. I. J. M., van Velzen, N., He, M., Lee, H., Noh, S. J., Rakovec, O., and Restrepo, P.: Advancing data assimilation in operational hydrologic forecasting: progresses, challenges, and emerging opportunities, Hydrology and Earth System Sciences, 16, 3863–3887, https://doi.org/10.5194/hess-16-3863-2012, 2012.

Lorenz, E. N.: Atmospheric predictability experiments with a large numerical model, Tellus, 34, 505–513, https://doi.org/10.1111/j.2153-3490.1982.tb01839.x, 1982.

Lorenz, E. N.: Predictability: A problem partly solved, in: Proc. Seminar on predictability, vol. 1, 1996.

Lorenz, E. N.: Designing chaotic models, Journal of the Atmospheric Sciences, 62, 1574–1587, https://doi.org/10.1175/JAS3430.1, 2005.

Lorenz, E. N. and Emanuel, K. A.: Optimal sites for supplementary weather observations: Simulation with a small model, Journal of the Atmospheric Sciences, 55, 399–414, https://doi.org/10.1175/1520-0469(1998)055<0399:OSFSWO>2.0.CO;2, 1998.

Luo, Y., Ogle, K., Tucker, C., Fei, S., Gao, C., LaDeau, S., Clark, J. S., and Schimel, D. S.: Ecological forecasting and data assimilation in a data-rich era, Ecological Applications, 21, 1429–1442, https://doi.org/10.1890/09-1275.1, 2011.

Mualem, Y.: A new model for predicting the hydraulic conductivity of unsaturated porous media, Water Resources Research, 12, 513–522, https://doi.org/10.1029/WR012i003p00513, 1976.

Nakano, S., Ueno, G., and Higuchi, T.: Merging particle filter for sequential data assimilation, Nonlinear Processes in Geophysics, 14, 395–408, https://doi.org/10.5194/npg-14-395-2007, 2007.

Niu, S., Luo, Y., Dietze, M. C., Keenan, T. F., Shi, Z., Li, J., and III, F. S. C.: The role of data assimilation in predictive ecology, Ecosphere, 5, art65, https://doi.org/10.1890/ES13-00273.1, 2014.

Poterjoy, J.: A localized particle filter for high-dimensional nonlinear systems, Monthly Weather Review, 144, 59–76, https://doi.org/10.1175/MWR-D-15-0163.1, 2016.

Rasmussen, J., Madsen, H., Jensen, K. H., and Refsgaard, J. C.: Data assimilation in integrated hydrological modeling using ensemble Kalman filtering: evaluating the effect of ensemble size and localization on filter performance, Hydrology and Earth System Sciences, 19, 2999–3013, https://doi.org/10.5194/hess-19-2999-2015, 2015.

Reichle, R. H.: Data assimilation methods in the Earth sciences, Advances in Water Resources, 31, 1411 – 1418, https://doi.org/10.1016/j.advwatres.2008.01.001, hydrologic Remote Sensing, 2008.

Ruiz, J. J., Pulido, M., and Miyoshi, T.: Estimating Model Parameters with Ensemble-Based Data Assimilation: A Review, Journal of the Meteorological Society of Japan. Ser. II, 91, 79–99, https://doi.org/10.2151/jmsj.2013-201, 2013.

Shi, Y., Davis, K. J., Zhang, F., Duffy, C. J., and Yu, X.: Parameter estimation of a physically based land surface hydrologic model using the ensemble Kalman filter: A synthetic experiment, Water Resources Research, 50, 706–724, https://doi.org/10.1002/2013WR014070, 2014.

Stammer, D., Balmaseda, M., Heimbach, P., Köhl, A., and Weaver, A.: Ocean data assimilation in support of climate applications: Status and perspectives, Annual Review of Marine Science, 8, 491–518, https://doi.org/10.1146/annurev-marine-122414-034113, 2016.

Van Genuchten, M. T.: A closed-form equation for predicting the hydraulic conductivity of unsaturated soils, Soil science society of America journal, 44, 892–898, https://doi.org/10.2136/sssaj1980.03615995004400050002x, 1980.

van Leeuwen, P. J.: Nonlinear data assimilation in geosciences: Fan extremely efficient particle filter, Quarterly Journal of the Royal Meteorological Society, 136, 1991–1999, https://doi.org/10.1002/qj.699, 2010.

van Leeuwen, P. J., Cheng, Y., and Reich, S.: Nonlinear data assimilation, vol. 2, Springer, https://doi.org/10.1007/978-3-319-18347-3, 2015.

Zhang, H., Hendricks Franssen, H.-J., Han, X., Vrugt, J. A., and Vereecken, H.: State and parameter estimation of two land surface models using the ensemble Kalman filter and the particle filter, Hydrology and Earth System Sciences, 21, 4927–4958, https://doi.org/10.5194/hess-21-4927-2017, 2017.

Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C., and Baklanov, A.: Real-time air quality forecasting, part II: State of the science, current research needs, and future prospects, Atmospheric Environment, 60, 656 – 676, https://doi.org/https://doi.org/10.1016/j.atmosenv.2012.02.041, 2012.