Assessing the influence of observations in ensemble-based data assimilation systems

Article

Published Version

Hu, G. ORCID: https://orcid.org/0000-0003-4305-3658, Dance, S. L. ORCID: https://orcid.org/0000-0003-1690-3338, Fowler, A. ORCID: https://orcid.org/0000-0003-3650-3948, Simonin, D. and Waller, J. (2025) Assessing the influence of observations in ensemble-based data assimilation systems. Journal of Advances in Modeling Earth Systems, 17 (11). e2024MS004809. ISSN 1942-2466 doi: 10.1029/2024MS004809 Available at https://centaur.reading.ac.uk/125280/

To link to this article DOI: http://dx.doi.org/10.1029/2024MS004809

Publisher: American Geophysical Union

JAMES | Journal of Advances in Modeling Earth Systems®

RESEARCH ARTICLE

10.1029/2024MS004809

Assessing the Influence of Observations in Ensemble-Based Data Assimilation Systems

Guannan Hu[1,2], Sarah L. Dance[1,2], Alison Fowler[1,2], David Simonin[3], and Joanne Waller[3]

[1]School of Mathematical, Physical and Computational Sciences, University of Reading, Reading, UK, [2]National Centre for Earth Observation (NCEO), University of Reading, Reading, UK, [3]Met Office, Reading, UK

Key Points:

- We propose new approaches for estimating the degrees of freedom for signal in ensemble-based data assimilation systems
- We present a general strategy for implementing these approaches in the presence of domain localization
- Numerical results show that the proposed approaches and strategy accurately estimate observation influence

Correspondence to:

G. Hu,
guannan.hu@reading.ac.uk

Citation:

Hu, G., Dance, S. L., Fowler, A., Simonin, D., & Waller, J. (2025). Assessing the influence of observations in ensemble-based data assimilation systems. *Journal of Advances in Modeling Earth Systems*, *17*, e2024MS004809. https://doi.org/10.1029/2024MS004809

Received 1 NOV 2024
Accepted 2 OCT 2025

Author Contributions:

Conceptualization: Guannan Hu, Sarah L. Dance, Alison Fowler, David Simonin, Joanne Waller
Funding acquisition: Sarah L. Dance, David Simonin
Project administration: Sarah L. Dance, David Simonin
Software: Guannan Hu
Supervision: Sarah L. Dance, Alison Fowler, David Simonin, Joanne Waller
Visualization: Guannan Hu
Writing – original draft: Guannan Hu
Writing – review & editing: Sarah L. Dance, Alison Fowler, Joanne Waller

Abstract

The skill of numerical weather forecasts strongly depends on the quality of the initial conditions (analyses), which are created by assimilating observations into previous short-range model forecasts. Therefore, it is important to carefully assess the influence of different observations on the analysis. The degrees of freedom for signal (DFS) is a useful metric for quantifying this influence. While DFS has long been used in variational data assimilation (DA) systems, its application in ensemble-based DA systems remains limited. In this study, we propose two novel approaches for estimating the DFS in ensemble-based systems. One approach uses the weighting vector calculated in ensemble transform Kalman filters, while the other uses the innovation vector and observation-space increment vector. We also propose a new strategy for implementing the DFS approaches in the presence of domain localization, which first estimates DFS locally and then aggregates the results to derive a global DFS value for each observation. Our numerical results show that the DFS per observation decreases as the localization radius increases. More generally, the proposed DFS approaches and implementation strategy have the potential to be used in practice to inform the optimization of observation networks and DA systems.

Plain Language Summary

Our daily weather forecasts rely on the use of weather observations (e.g., those from weather stations, satellites and radar). A better forecast can be achieved by improvements to the observation network and better use of the observations. To this end, we need to understand which types of observations are more valuable and whether they are being used optimally. Therefore, we propose new approaches for quantifying the value of different observations for weather forecasting. These approaches are specifically used in ensemble forecasting systems, which are widely used to predict high-impact weather events such as heavy precipitation.

1. Introduction

Weather observations are used to generate initial conditions (analyses) for numerical weather prediction (NWP) through a process called data assimilation (DA) (Ballard et al., 2016; Brousseau et al., 2012; Fischer et al., 2005; Lorenc et al., 2000; Rabier et al., 2000; Rawlins et al., 2007). They contribute greatly to improving the accuracy of weather forecasting (e.g., Bauer et al., 2015). Different observation types (e.g., satellite data, radar data, and aircraft data) provide different information on a variety of model variables. In addition, they have different spatial distributions, frequencies and error statistics and are often associated with different observation operators. Therefore, their influence on the analysis should be quantified so that we can understand the relative importance of different observation types for constraining the analysis.

An assessment of observation influence provides evidence to inform further development of current observation networks or the design of future ones. In addition, it can help diagnose problems with the assimilation of observations (Ota et al., 2013; Stiller, 2022). Some observation types may not be assimilated optimally due to reasons such as incorrectly specified observation error statistics (e.g., Fowler et al., 2020). Another purpose of quantifying the influence of observations is to assess changes to the DA system, for example, tuning the observation error covariance matrix (Chapnik et al., 2006; Desroziers & Ivanov, 2001) and localization radius (Diefenbach et al., 2022; Vural et al., 2024). Mostly, we are interested in the influence of a given type of observation, but sometimes it is also useful to separate the influence of each observation within a group in different ways (e.g., time of day, geographical area, etc.). This allows us to identify the dependence of the influence of observations on underlying meteorology.

The degrees of freedom for signal (DFS) is an information content measure used to quantify how much information the analysis has extracted from the observations (Rodgers, 1998, 2000). The DFS has been used in operational variational DA systems to select satellite channels (e.g., Collard, 2007) and tune observation error covariance matrices (Chapnik et al., 2006; Desroziers & Ivanov, 2001). How to estimate the DFS depends on how the assimilation system is implemented. In variational DA systems, there are a few different approaches for estimating it. Fisher (2003) has shown that the DFS can be expressed as the trace of a function of the Hessian matrix, which can be efficiently estimated using the algorithm of Bai et al. (1996). An alternative approach uses a randomized trace estimation method (Desroziers, Brousseau, & Chapnik, 2005; Fisher, 2003; Wahba et al., 1995). In addition, Cardinali et al. (2004) estimated the DFS using a low-rank approximation of the analysis error covariance matrix. More recently, Fowler et al. (2020) advocated using assimilation residuals in observation space to simultaneously estimate the theoretical DFS (i.e. consistent with the assumptions made in the assimilation) and the actual DFS (that accounts for violations of these assumptions). This makes the DFS useful for assessing not only the influence of the observations but also whether this influence is optimal.

Ensemble Kalman filters represent another category of operational DA systems (Carrera et al., 2015; Schraff et al., 2016). Due to advantages such as the use of flow-dependent background error statistics, the ability to generate initial conditions for ensemble forecasts, computational efficiency, and relative ease of implementation, ensemble Kalman filters are gaining popularity and feasibility for convection-permitting NWP (e.g., Hu et al., 2023), among other applications. Estimating the DFS in such systems is different from estimating it in variational DA systems. On the one hand, in ensemble Kalman filters, the Kalman gain can be explicitly formed using background or analysis ensemble perturbations (Liu et al., 2009), making the estimation of the DFS potentially easier (Hotta & Ota, 2021). On the other hand, the ensemble estimate of the background error covariance matrix is flow-dependent and contains ensemble sampling error, leading to spatial and temporal variations in the influence of the same type of observations. This means that it is more difficult to obtain statistically significant estimates of DFS. In practice, we need to decide on which time period and in which region to average the DFS.

Domain localization provides a method for the local implementation of ensemble Kalman filters. It divides a large global DA problem into many smaller independent local DA problems (Greybush et al., 2011; Janjić et al., 2011; Schraff et al., 2016). Each local analysis process calculates the analysis state at one or several model grid points using only observations within a certain distance from those grid points. Consequently, each local analysis uses a local background ensemble perturbation matrix, a local background ensemble perturbation matrix in observation space, and a local observation error covariance matrix. These local matrices are composed of selected elements of their global counterparts (for a detailed example see Section 6.1). As a result, the local Kalman gains differ across local analyses and do not correspond to rows of a single global Kalman gain calculated using global error covariance matrices. While domain localization brings computational benefits to ensemble Kalman filters, it complicates the estimation of the DFS (Hotta & Ota, 2021).

To the best of our knowledge, the literature on the use of the DFS in ensemble-based DA systems remains limited, and the answers to the following questions are not yet clear.

- Which approaches can be used to estimate the DFS in ensemble-based DA systems?
- How can the DFS be reliably and efficiently estimated in the presence of domain localization?

In this work, we revisit existing state-of-the-art approaches (Fowler et al., 2020; Hotta & Ota, 2021) and propose two novel approaches for estimating the DFS in various ensemble Kalman filter frameworks. In addition, we propose a new strategy for implementing the DFS approaches in the presence of domain localization and compare it with the strategy proposed by Hotta and Ota (2021).

This paper is organized as follows. In the next section, we introduce some notation and basic concepts of ensemble-based DA. In Section 3, we introduce the definition of the DFS, including both the actual and theoretical DFS formulations. In Section 4, we propose two novel approaches for estimating the DFS and revisit several existing approaches. In Section 5, we propose a new strategy for implementing the DFS approaches in the presence of domain localization. In Section 6, we conduct numerical experiments to study the uncertainty in the estimates of the actual DFS, the sensitivity of DFS estimates to ensemble size, the effect of domain localization on DFS estimation, and the spatial variability of the DFS for individual observations. Finally, in Section 7, we summarize the advantages and disadvantages of the different DFS approaches and our new strategy for handling domain localization, and address practical considerations for applying these approaches in operational environments.

2. Mathematical Concepts and Notation for Data Assimilation

In this section, we introduce some basic concepts for ensemble-based DA (following standard notation; e.g., Livings et al., 2008). Let $\mathbf{x} \in \mathbb{R}^n$ be a model state vector. The ensemble mean of $\mathbf{x}$ is

$$\overline{\mathbf{x}} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{x}_k,$$

where k denotes the kth ensemble member and K the ensemble size. The ensemble perturbation matrix is the $n \times K$ matrix defined by

$$\mathbf{X} = [\mathbf{x}_1 - \overline{\mathbf{x}} \quad \mathbf{x}_2 - \overline{\mathbf{x}} \quad \dots \quad \mathbf{x}_K - \overline{\mathbf{x}}].$$

The corresponding ensemble covariance matrix is the $n \times n$ matrix given by

$$\mathbf{P} = \frac{1}{K-1}\mathbf{X}\mathbf{X}^\top, \tag{1}$$

where the superscript $\top$ denotes the transpose of a matrix. This ensemble covariance matrix is subject to *ensemble sampling error*, as it is computed from a finite ensemble of size K.

Let $h : \mathbb{R}^n \to \mathbb{R}^m$ denote a nonlinear observation operator. The ensemble mean of the model state vector in observation space is

$$\overline{h(\mathbf{x})} = \frac{1}{K}\sum_{k=1}^{K}h(\mathbf{x}_k).$$

Let $\mathbf{y} \in \mathbb{R}^m$ be an observation vector. The ensemble mean of the innovation vector is

$$\overline{\mathbf{d}} = \mathbf{y} - \overline{h(\mathbf{x}_\mathrm{b})}, \tag{2}$$

where $\mathbf{x}_\mathrm{b}$ is the background model state vector and we omit the ensemble index k. Let $\mathbf{H} \in \mathbb{R}^{m\times n}$ be the linearized observation operator. The ensemble perturbation matrix in observation space is

$$\mathbf{Y} = \mathbf{H}\mathbf{X}, \tag{3}$$

which is an $m \times K$ matrix. It should be noted that the definition of the matrix $\mathbf{Y}$ can be extended to nonlinear observation operators by computing each column as $h(\mathbf{x}_k) - \overline{h(\mathbf{x})}$. Let $\mathbf{R} \in \mathbb{R}^{m\times m}$ be the assumed observation error covariance matrix. The term "assumed" indicates that it is the matrix used in the DA and may not represent the true error statistics. The corresponding *assumed* innovation covariance matrix is then

$$\mathbf{D} = \mathbf{H}\mathbf{P}_\mathrm{b}\mathbf{H}^\top + \mathbf{R}, \tag{4}$$

where $\mathbf{P}_\mathrm{b} \in \mathbb{R}^{n\times n}$ is the ensemble background error covariance matrix defined by Equation 1 (Todling et al., 1998). The matrix $\mathbf{D}$ is considered to follow an unbiased Gaussian distribution. The Kalman gain is the $n \times m$ matrix defined by

$$\mathbf{K} = \mathbf{P}_\mathrm{b}\mathbf{H}^\top\mathbf{D}^{-1}, \tag{5}$$

which is equivalent to

$$\mathbf{K} = \mathbf{P}_\mathrm{a}\mathbf{H}^\top\mathbf{R}^{-1}, \tag{6}$$

where $\mathbf{P}_\mathrm{a} \in \mathbb{R}^{n\times n}$ is the ensemble analysis error covariance matrix (Kalnay et al., 2012). Using the Kalman gain, the ensemble mean of the analysis state vector can be calculated by

$$\overline{\mathbf{x}}_\mathrm{a} = \overline{\mathbf{x}}_\mathrm{b} + \mathbf{K}\overline{\mathbf{d}}, \tag{7}$$

which corresponds to the best linear unbiased estimator (Nichols, 2010).

To account for inaccuracies in the assumed error statistics, we define the *true* innovation covariance matrix as

$$\mathbb{E}\left[\overline{\mathbf{d}}\,\overline{\mathbf{d}}^\top\right] \approx \mathbf{H}\mathbf{B}_\mathrm{t}\mathbf{H} + \mathbf{R}_\mathrm{t}, \tag{8}$$

where $\mathbb{E}[\cdot]$ denotes the statistical expectation, and $\mathbf{R}_\mathrm{t} \in \mathbb{R}^{m\times m}$ and $\mathbf{B}_\mathrm{t} \in \mathbb{R}^{n\times n}$ are the true observation and background error covariance matrices, respectively. If the background ensemble perturbations are representative of the true background error statistics, then $\mathbf{B}_\mathrm{t} = \mathbb{E}[\mathbf{P}_\mathrm{b}]$. The derivation of Equation 8 is provided in Appendix A, following Desroziers, Berre, et al. (2005). The true innovation covariance matrix can be empirically estimated as

$$\mathbb{E}\left[\overline{\mathbf{d}}\,\overline{\mathbf{d}}^\top\right] \approx \frac{1}{N}\sum_{j=1}^{N}\overline{\mathbf{d}}_j\,\overline{\mathbf{d}}_j^\top, \tag{9}$$

where j is the sample index and N the total number of samples. The expectation is taken over the probability density function of actual innovation samples, which are expected to follow an unbiased Gaussian distribution when the observation operator is linear. The estimated covariance matrix is subject to *innovation sampling error*, which arises from two main sources: (a) sampling of observations and (b) sampling of the background mean (Equation 2). To avoid biased estimates, the sampled innovation vector $\overline{\mathbf{d}}$ should reflect consistent innovation statistics. Therefore, innovation samples should be collected from different analysis times and/or regions where observation and background error statistics are assumed to be consistent or sufficiently similar. Otherwise, the resulting estimate would represent an average over a mixture of different innovation covariance matrices.

3. Degrees of Freedom for Signal

The DFS is named after the idea that model state space can be spanned by p orthogonal vectors so that it can vary statistically independently in p directions ("degrees of freedom"; Section 2.4 of Rodgers, 2000). If the observations constrain the uncertainty well in one direction, it can be considered to represent one "degree of freedom for signal." The unconstrained directions, on the other hand, represent the "degrees of freedom for noise". Hence, a larger DFS means that the observations have a larger influence on the analysis. By definition, the smallest value of the DFS is zero, and the largest value is the smaller of the number of observations (m) and the dimension of the model (n). However, in ensemble-based DA, the background error covariance matrix is estimated using the ensemble perturbation matrix (Equation 1) and thus has a maximum rank of $K - 1$. In this case, the value of the DFS is bounded by $K - 1$, much smaller than m or n (Hotta & Ota, 2021). If background error covariance localization is applied, then the upper bound can increase to $K \cdot \mathrm{rank}(\mathbf{L}_\mathrm{nn}) - 1$, where $\mathbf{L}_\mathrm{nn} \in \mathbb{R}^{n\times n}$ is a model-space localization matrix (Hotta & Ota, 2021). In ensemble-based DA, we may think of the DFS as measuring how many of the directions spanned by the ensemble members are constrained by the observations.

For a given background error covariance matrix, the DFS can be expressed as

$$\mathrm{DFS} = \mathbb{E}\left[(\overline{\mathbf{x}}_\mathrm{a} - \overline{\mathbf{x}}_\mathrm{b})^\top\mathbf{P}_\mathrm{b}^+(\overline{\mathbf{x}}_\mathrm{a} - \overline{\mathbf{x}}_\mathrm{b})\right], \tag{10}$$

following Section 2.4 of Rodgers (2000). Here, we use a pseudoinverse $\mathbf{P}_\mathrm{b}^+$ that satisfies $\mathbf{P}_\mathrm{b}\mathbf{P}_\mathrm{b}^+\mathbf{P}_\mathrm{b} = \mathbf{P}_\mathrm{b}$ (Chapter 5 of Golub & Van Loan, 1996), as the matrix $\mathbf{P}_\mathrm{b}$ may not be invertible. Since this expression can later be reformulated in terms of $\mathbb{E}\left[\overline{\mathbf{d}}\,\overline{\mathbf{d}}^\top\right]$, we can consider that the expectation is taken over an unbiased Gaussian distribution, as in Equation 9. A direct interpretation of Equation 10 is that a larger DFS implies a larger update to the background state vector due to the assimilation of observations. Moreover, since the analysis is the best linear unbiased estimate, a larger DFS also indicates a larger reduction in the variance of the analysis error (with respect to the variance of the background error) and, hence, a more accurate analysis.

The equation for the actual DFS can be obtained by substituting Equations 7 and 5 into Equation 10. As derived in Appendix B, we obtain

$$\mathrm{DFS}_{\mathrm{act}} = \mathrm{tr}\left(\mathbb{E}\left[\overline{\mathbf{d}}\,\overline{\mathbf{d}}^{\top}\right]\mathbf{D}^{-1}\mathbf{H}\mathbf{K}\right), \tag{11}$$

which measures the *actual* influence of observations, taking into account the difference between assumed and true innovation covariance matrices (Fowler et al., 2020). It should be noted that $\mathrm{DFS}_{\mathrm{act}}$ is derived under the assumption of stationary background and observation error statistics. In ensemble-based DA, the background error statistics are flow-dependent, and samples of $\overline{\mathbf{d}}$ should be carefully selected to ensure this assumption is reasonably satisfied.

If the two innovation covariance matrices are identical, then Equation 11 becomes

$$\mathrm{DFS}_{\mathrm{theo}} = \mathrm{tr}(\mathbf{H}\mathbf{K}), \tag{12}$$

which gives the *theoretical* value of the DFS. The theoretical DFS may also be seen as the trace of the derivative of analysis in observation space with respect to observations (Cardinali et al., 2004; Chapnik et al., 2006). In practice, if the background and observation error covariance matrices used in DA represent the true error statistics well and the observation operator is close to linear, then the theoretical and actual values of the DFS are expected to be similar. The difference between the theoretical and actual DFS can be used to evaluate deviations from these assumptions. When the deviations are large, the actual DFS provides an accurate estimate of Equation 10, properly accounting for these deviations.

The theoretical DFS is determined by the background and observation error statistics and the observation operator. In DA systems with no cycling, this means that it is independent of the numerical values of the observations. In other words, observations with different values can have the same theoretical DFS as long as the error statistics are the same and the observation operators are the same. However, the numerical value of observations starts to play a role in cycling systems because it affects the forecast ensemble used to estimate the background error statistics in the next cycle. Compared to the theoretical DFS, the actual DFS also involves the true innovation covariance, which is estimated using a sample of the vector $\overline{\mathbf{d}}$ (Equation 9). Therefore, the exact value of the observations affects the estimate of the actual DFS.

4. Methodology for Estimating DFS

In this section, we describe how to efficiently estimate the theoretical and actual DFS in ensemble-based DA systems. The equations used to define the DFS may not be directly usable in practice. The estimation of the theoretical DFS (Equation 12) requires an explicit formulation of the Kalman gain acted on by a linear observation operator or some approximation of the matrix product, **HK**. The estimation of the actual DFS (Equation 11) is more complicated because we also need an estimate of the true innovation covariance matrix.

4.1. Weighting-Vector-Based Approach

We first introduce a novel approach for estimating the actual DFS in ensemble transform Kalman filters (ETKF; Bishop et al., 2001; Hunt et al., 2007), where the ensemble mean of the analysis state vector is calculated by

$$\overline{\mathbf{x}}_{\mathrm{a}} = \overline{\mathbf{x}}_{\mathrm{b}} + \mathbf{X}_{\mathrm{b}}\mathbf{w}, \tag{13}$$

with $\mathbf{w} \in \mathbb{R}^{K}$ being a weighting vector. Using this vector, the actual DFS can be estimated by

$$\mathrm{DFS}_{\mathrm{act,w}} = \frac{K-1}{N}\sum_{j=1}^{N}\mathbf{w}_j^{\top}\mathbf{w}_j. \tag{14}$$

Ideally, samples of the vector **w** should be collected under stationary background error statistics. However, this is difficult in practice due to the flow-dependent nature of background error statistics. Therefore, we recommend collecting samples from assimilation times and/or regions where the background error statistics are supposed to be sufficiently similar (e.g., during a specific weather event).

The $\mathrm{DFS}_{\mathrm{act,w}}$ approach is computationally inexpensive, as the vector **w** is an intermediate product calculated in the ETKF algorithm. Since the vector **w** is not typically an output of operational systems, we may need to apply Equation 14 alongside the assimilation rather than as a post-processing method.

It is not possible to separate the influence of a subset of observations directly from Equation 14 as the vector **w** is in ensemble space (of size K). We define $\mathbf{\Pi}_i$ as a projection operator that selects the vector elements corresponding to the subset i of observations, for example, $\mathbf{\Pi}_i\,\mathbf{y}$ (Desroziers, Brousseau, & Chapnik, 2005; Fowler et al., 2020; Stiller, 2022). The operator $\mathbf{\Pi}_i$ is a row vector or a matrix that contains only elements of zero and one. We further define the matrix $\mathbf{S}_i = \mathbf{\Pi}_i^{\top}\mathbf{\Pi}_i$. Then, the actual DFS for the subset i of observations can be calculated by

$$\mathrm{DFS}_{\mathrm{act,w},i} = \frac{1}{N}\sum_{j=1}^{N}\left(\overline{\mathbf{d}}_j - \mathbf{Y}_{\mathrm{b},j}\mathbf{w}_j\right)^{\top}\mathbf{R}^{-1}\mathbf{S}_i\,\mathbf{Y}_{\mathrm{b},j}\mathbf{w}_j, \tag{15}$$

where $\mathbf{Y}_{\mathrm{b},j}$ denotes the jth sample of the matrix $\mathbf{Y}_{\mathrm{b}}$ (Equation 3). The vector $\overline{\mathbf{d}}$ and the matrix $\mathbf{Y}_{\mathrm{b}}$ are usually readily available in ETKF systems. They are sampled alongside the vector **w**. For the derivation of the weighting-vector-based approach, see Appendix C.

4.2. Ensemble Perturbation Approaches

This type of approach has been used by Hotta and Ota (2021) to estimate the theoretical DFS. The idea is straightforward: we can use either background or analysis ensemble perturbations to explicitly form the Kalman gain (Liu et al., 2009) and then compute the matrix **HK**. Using the Kalman gain given by Equation 5, along with Equations 1 and 3, the theoretical DFS can be estimated by

$$\mathrm{DFS}_{\mathrm{theo},\mathrm{Y_b}} = \mathrm{tr}\left(\mathbf{Y}_{\mathrm{b}}\,\mathbf{Y}_{\mathrm{b}}^{\top}\left(\mathbf{Y}_{\mathrm{b}}\,\mathbf{Y}_{\mathrm{b}}^{\top} + (K-1)\mathbf{R}\right)^{-1}\right), \tag{16}$$

and the contribution of the subset i of observations to the total DFS is

$$\mathrm{DFS}_{\mathrm{theo},\mathrm{Y_b},i} = \mathrm{tr}\left(\mathbf{Y}_{\mathrm{b}}\,\mathbf{Y}_{\mathrm{b}}^{\top}\left(\mathbf{Y}_{\mathrm{b}}\,\mathbf{Y}_{\mathrm{b}}^{\top} + (K-1)\mathbf{R}\right)^{-1}\mathbf{S}_i\right). \tag{17}$$

Alternatively, using the expression for the Kalman gain in Equation 6, the theoretical DFS can be estimated from the analysis ensemble perturbation as

$$\mathrm{DFS}_{\mathrm{theo},\mathrm{Y_a}} = \frac{1}{K-1}\mathrm{tr}\left(\mathbf{Y}_{\mathrm{a}}\,\mathbf{Y}_{\mathrm{a}}^{\top}\mathbf{R}^{-1}\right), \tag{18}$$

with the contribution of observation subset i given by

$$\mathrm{DFS}_{\mathrm{theo},\mathrm{Y_a},i} = \frac{1}{K-1}\mathrm{tr}\left(\mathbf{Y}_{\mathrm{a}}\,\mathbf{Y}_{\mathrm{a}}^{\top}\mathbf{R}^{-1}\mathbf{S}_i\right). \tag{19}$$

The $\mathrm{DFS}_{\mathrm{theo},\mathrm{Y_a}}$ formulation is computationally cheaper than $\mathrm{DFS}_{\mathrm{theo},\mathrm{Y_b}}$ if the matrix $\mathbf{Y}_{\mathrm{a}}$ is readily available or can be efficiently computed.

In ETKF systems, assuming a linear observation operator, the matrix $\mathbf{Y}_a$ can be calculated as

$$\mathbf{Y}_a = \mathbf{Y}_b \mathbf{W}, \tag{20}$$

where

$$\mathbf{W} = \left((K-1)\widetilde{\mathbf{P}}_a\right)^{1/2}$$

is a transformation matrix obtained during the assimilation process (e.g., Hunt et al., 2007) and the matrix $\widetilde{\mathbf{P}}_a$ is given by Equation C3.

In stochastic ensemble Kalman filters, the matrix $\mathbf{Y}_a$ can be obtained by subtracting the ensemble mean from each member of the analysis ensemble in observation space (calculated using a nonlinear observation operator). This is not desirable when domain localization is used, as the observation operator is often non-local (see Section 5). Alternatively, we may use a linearized observation operator to compute the matrix $\mathbf{Y}_a$ as $\mathbf{HX}_a$. A potential issue with this is that the matrix $\mathbf{X}_a$ is typically inflated in operational DA systems (Whitaker & Hamill, 2012; Zhang et al., 2004), and using an artificially inflated $\mathbf{X}_a$ can distort the estimation of the DFS. Therefore, the application of the $\mathrm{DFS}_{\mathrm{theo},\mathrm{Y}_a}$ approach in stochastic filters is less straightforward than in ETKF systems, where Equation 20 is applicable.

4.3. Innovation-Based Approaches

The innovation-based approaches were originally proposed for variational DA systems (Fowler et al., 2020; Lupu et al., 2011). They provide estimates of both the theoretical and actual DFS. We describe how to use them in ensemble-based DA systems and provide a new formulation for estimating the theoretical DFS. We may consider that the idea of the innovation-based approaches is to form the Kalman gain using the innovation, residual and observation-space increment vectors. The innovation vector is defined in Equation 2 as it is required for DA. The residual and observation-space increment vectors are defined as

$$\mathbf{r} = \mathbf{y} - h(\overline{\mathbf{x}}_a), \tag{21}$$

and

$$\mathbf{v} = h(\overline{\mathbf{x}}_a) - h(\overline{\mathbf{x}}_b). \tag{22}$$

Depending on the DA system, these two vectors may also be calculated using the ensemble mean of $h(\mathbf{x}_a)$ and $h(\mathbf{x}_b)$ or the control member (see discussion in Appendix D). We further normalize the vectors by the square root of the observation error precision matrix, which can be expressed as

$$\begin{aligned} \hat{\mathbf{d}} &= \mathbf{R}^{-1/2}\overline{\mathbf{d}}, \\ \hat{\mathbf{r}} &= \mathbf{R}^{-1/2}\mathbf{r}, \\ \hat{\mathbf{v}} &= \mathbf{R}^{-1/2}\mathbf{v}. \end{aligned} \tag{23}$$

The purpose of the normalization is explained in Section 4.3.1. For diagonal observation error covariance matrices, the three vectors are simply scaled by the observation error standard deviation.

4.3.1. Fowler et al. (2020) Approach

The actual DFS (Equation 11) can be estimated by

$$\mathrm{DFS}_{\mathrm{act,d}} = \frac{1}{N}\sum_{j=1}^{N} \hat{\mathbf{r}}_j^{\top}\hat{\mathbf{v}}_j, \tag{24}$$

and the influence of the subset i of observations is

$$\mathrm{DFS}_{\mathrm{act,d},i} = \frac{1}{N}\sum_{j=1}^{N} \hat{\mathbf{r}}_j^{\top}\mathbf{S}_i\hat{\mathbf{v}}_j. \tag{25}$$

The theoretical DFS (Equation 12) can be estimated by

$$\mathrm{DFS}_{\mathrm{theo,d}} = \mathrm{tr}\left(\sum_{j=1}^{N}\hat{\mathbf{v}}_j\hat{\mathbf{r}}_j^{\top}\left(\sum_{j=1}^{N}\hat{\mathbf{d}}_j\hat{\mathbf{r}}_j^{\top}\right)^{-1}\right), \tag{26}$$

and

$$\mathrm{DFS}_{\mathrm{theo,d},i} = \mathrm{tr}\left(\sum_{j=1}^{N}\hat{\mathbf{v}}_j\hat{\mathbf{r}}_j^{\top}\left(\sum_{j=1}^{N}\hat{\mathbf{d}}_j\hat{\mathbf{r}}_j^{\top}\right)^{-1}\mathbf{S}_i\right). \tag{27}$$

For the derivation of these equations, see Fowler et al. (2020).

In the above equations, samples of the vectors $\hat{\mathbf{d}}$, $\hat{\mathbf{r}}$ and $\hat{\mathbf{v}}$ should be collected at different assimilation times and/or regions where the background error statistics are similar. In general, how the sample is selected for averaging should depend on the practical application, such as the type of observation or the purpose of the observation influence assessment.

As shown by Equation 26, matrix inversion is required for estimating the theoretical DFS. This may lead to large numerical errors if the matrix is ill-conditioned. Without the normalization given by Equation 23, we need to invert the matrix $\sum_{j=1}^{N}\mathbf{d}_j\mathbf{r}_j^{\top}$, while with the normalization, we need to invert the matrix $\sum_{j=1}^{N}\hat{\mathbf{d}}_j\hat{\mathbf{r}}_j^{\top}$, which is expected to be better conditioned (Fowler et al., 2020; Tabeart et al., 2018).

4.3.2. New Alternative Approach

In addition to the original approach, $\mathrm{DFS}_{\mathrm{theo,d}}$, we propose a new alternative approach to estimate the theoretical DFS (for a derivation, see Appendix E). The equations for the alternative approach are obtained by replacing the residual vector in the original equations (Equations 26 and 27) with the innovation vector. The new equations are

$$\mathrm{DFS}_{\mathrm{theo,d,alt}} = \mathrm{tr}\left(\sum_{j=1}^{N}\hat{\mathbf{v}}_j\hat{\mathbf{d}}_j^{\top}\left(\sum_{j=1}^{N}\hat{\mathbf{d}}_j\hat{\mathbf{d}}_j^{\top}\right)^{-1}\right) \tag{28}$$

for all observations and

$$\mathrm{DFS}_{\mathrm{theo,d,alt},i} = \mathrm{tr}\left(\sum_{j=1}^{N}\hat{\mathbf{v}}_j\hat{\mathbf{d}}_j^{\top}\left(\sum_{j=1}^{N}\hat{\mathbf{d}}_j\hat{\mathbf{d}}_j^{\top}\right)^{-1}\mathbf{S}_i\right) \tag{29}$$

for the subset i of the observations.

The original and alternative approaches each have their own advantages. An advantage of the alternative approach is that it simplifies the computation by using only the innovation and observation-space increment vectors. For the original approach, the better the assimilation is (i.e., the closer the assumed innovation covariance matrix is to the true innovation covariance matrix), the closer the matrix to be inverted $\left(\sum_{j=1}^{N}\hat{\mathbf{d}}_j\hat{\mathbf{r}}_j^{\top}\right)$ will be to the identity matrix. Therefore, we may assume this matrix to be diagonal and use only the diagonal elements to estimate the theoretical DFS (Fowler et al., 2020). This reduces the computational cost and avoids the numerical error that may be caused by matrix inversion. For the alternative approach, the matrix to be inverted is the normalized innovation covariance matrix $\left(\sum_{j=1}^{N} \hat{\mathbf{d}}_j \hat{\mathbf{d}}_j^\top\right)$. In general cases, this matrix is unlikely to be a diagonal matrix because it includes the background error covariances (A special case is when the matrix $\mathbf{H}\mathbf{P}_\mathrm{b}\mathbf{H}^\top$ is similar to the matrix $\mathbf{R}$).

4.4. Comparison Between Different DFS Approaches

In this section, we compare the six different DFS approaches introduced in the last section. Among them, the $\mathrm{DFS}_{\mathrm{act,w}}$ approach (Equation 14) and the $\mathrm{DFS}_{\mathrm{act,d}}$ (Equation 24) approach estimate the actual DFS, while the $\mathrm{DFS}_{\mathrm{theo,Y_b}}$ approach (Equation 16), the $\mathrm{DFS}_{\mathrm{theo,Y_a}}$ approach (Equation 18), the $\mathrm{DFS}_{\mathrm{theo,d}}$ approach (Equation 26) and the $\mathrm{DFS}_{\mathrm{theo,d,alt}}$ approach (Equation 28) estimate the theoretical DFS.

We begin by comparing the two approaches for estimating the actual DFS. The $\mathrm{DFS}_{\mathrm{act,w}}$ approach can be rewritten as

$$\mathrm{DFS}_{\mathrm{act,w}} = \frac{1}{(K-1)N} \sum_{j=1}^{N} \bar{\mathbf{d}}_j^\top \widetilde{\mathbf{D}}^{-1} \mathbf{Y}_\mathrm{b}\, \mathbf{Y}_\mathrm{b}^\top \widetilde{\mathbf{D}}^{-1} \bar{\mathbf{d}}_j,$$

where $\widetilde{\mathbf{D}}$ is the assumed innovation covariance matrix but calculated using a nonlinear observation operator (for a derivation, see Appendix C). The $\mathrm{DFS}_{\mathrm{act,d}}$ approach uses normalized residual and observation-space increment vectors, both of which can be expressed in terms of the innovation vector as follows

$$\hat{\mathbf{r}} \approx \mathbf{R}^{-1/2} (\mathbf{I} - \mathbf{H}\mathbf{K})\, \bar{\mathbf{d}} \tag{30}$$

and

$$\hat{\mathbf{v}} \approx \mathbf{R}^{-1/2} \mathbf{H}\mathbf{K} \bar{\mathbf{d}}, \tag{31}$$

where the approximation error is caused by the linearization of the nonlinear observation operator. Substituting Equations 30 and 31 into Equation 24, we obtain the following approximation

$$\mathrm{DFS}_{\mathrm{act,d}} \approx \frac{1}{(K-1)N} \sum_{j=1}^{N} \bar{\mathbf{d}}_j^\top \widetilde{\mathbf{D}}^{-1} \mathbf{H}\mathbf{X}_\mathrm{b}\, \mathbf{Y}_\mathrm{b}^\top \widetilde{\mathbf{D}}^{-1} \bar{\mathbf{d}}_j.$$

When the observation operator is linear, this approximation becomes exact and we have $\mathbf{Y}_\mathrm{b} = \mathbf{H}\mathbf{X}_\mathrm{b}$. Therefore, the two actual DFS approaches are equivalent in the case of a linear observation operator.

We now demonstrate the equivalence between the theoretical DFS approaches. We substitute Equation 30 into Equation 26, which gives

$$\begin{aligned} \mathrm{DFS}_{\mathrm{theo,d}} &\approx \mathrm{tr}\left(\sum_{j=1}^{N} \hat{\mathbf{v}}_j \bar{\mathbf{d}}_j^\top (\mathbf{I} - \mathbf{H}\mathbf{K})^\top \mathbf{R}^{-1/2} \left(\sum_{j=1}^{N} \hat{\mathbf{d}}_j \bar{\mathbf{d}}_j^\top (\mathbf{I} - \mathbf{H}\mathbf{K})^\top \mathbf{R}^{-1/2} \right)^{-1} \right) \\ &= \mathrm{tr}\left(\sum_{j=1}^{N} \hat{\mathbf{v}}_j \bar{\mathbf{d}}_j^\top \left(\sum_{j=1}^{N} \hat{\mathbf{d}}_j \bar{\mathbf{d}}_j^\top \right)^{-1} \right). \end{aligned}$$

By further applying Equation 23, we find that the $\mathrm{DFS}_{\mathrm{theo,d}}$ approach is equivalent to the $\mathrm{DFS}_{\mathrm{theo,d,alt}}$ approach. Using Equation 31, we can show that the two approaches approximate $\mathrm{tr}(\mathbf{HK})$ when the observation operator is nonlinear, and yield exactly $\mathrm{tr}(\mathbf{HK})$ when the observation operator is linear. Similarly, the $\mathrm{DFS}_{\mathrm{theo,Y_b}}$ and $\mathrm{DFS}_{\mathrm{theo,Y_a}}$ approaches also produce results equal to $\mathrm{tr}(\mathbf{HK})$ in the linear case. Therefore, all four theoretical DFS approaches are equivalent when the observation operator is linear.

When the observation operator is nonlinear, the $\mathrm{DFS}_{\mathrm{act,w}}$, $\mathrm{DFS}_{\mathrm{theo,Y_a}}$ and $\mathrm{DFS}_{\mathrm{theo,Y_b}}$ approaches can use either the nonlinear operator or its linearized version, consistent with the implementation in the DA system. Differences in

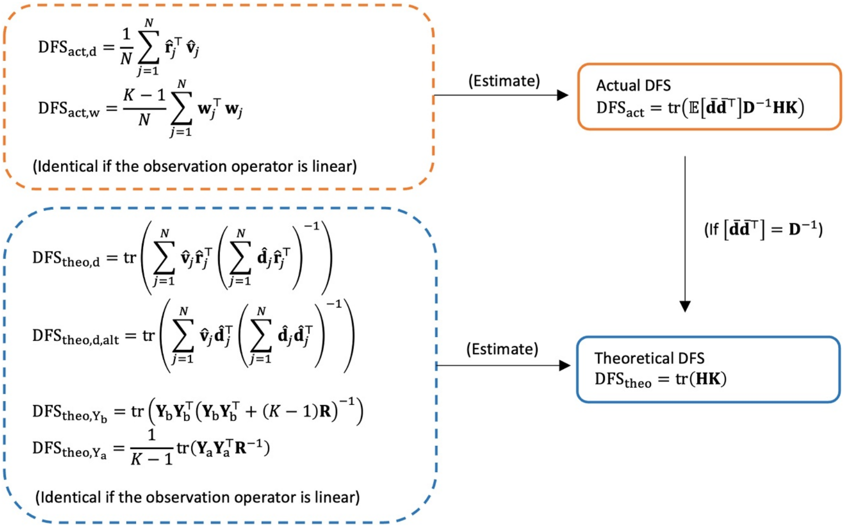
$\mathrm{DFS}_{\mathrm{act,d}} = \frac{1}{N}\sum_{j=1}^{N} \hat{\mathbf{r}}_j^{\mathrm{T}} \hat{\mathbf{v}}_j$
$\mathrm{DFS}_{\mathrm{act,w}} = \frac{K-1}{N}\sum_{j=1}^{N} \mathbf{w}_j^{\mathrm{T}} \mathbf{w}_j$
(Identical if the observation operator is linear)
(Estimate)
Actual DFS
$\mathrm{DFS}_{\mathrm{act}} = \mathrm{tr}\left(\mathbb{E}\left[\bar{\mathbf{d}}\bar{\mathbf{d}}^{\mathrm{T}}\right]\mathbf{D}^{-1}\mathbf{HK}\right)$
(If $\left[\bar{\mathbf{d}}\bar{\mathbf{d}}^{\mathrm{T}}\right] = \mathbf{D}^{-1}$)
$\mathrm{DFS}_{\mathrm{theo,d}} = \mathrm{tr}\left(\sum_{j=1}^{N} \hat{\mathbf{v}}_j \hat{\mathbf{r}}_j^{\mathrm{T}} \left(\sum_{j=1}^{N} \hat{\mathbf{d}}_j \hat{\mathbf{r}}_j^{\mathrm{T}}\right)^{-1}\right)$
$\mathrm{DFS}_{\mathrm{theo,d,alt}} = \mathrm{tr}\left(\sum_{j=1}^{N} \hat{\mathbf{v}}_j \hat{\mathbf{d}}_j^{\mathrm{T}} \left(\sum_{j=1}^{N} \hat{\mathbf{d}}_j \hat{\mathbf{d}}_j^{\mathrm{T}}\right)^{-1}\right)$
$\mathrm{DFS}_{\mathrm{theo,Y_b}} = \mathrm{tr}\left(\mathbf{Y}_{\mathrm{b}}\mathbf{Y}_{\mathrm{b}}^{\mathrm{T}}\left(\mathbf{Y}_{\mathrm{b}}\mathbf{Y}_{\mathrm{b}}^{\mathrm{T}} + (K-1)\mathbf{R}\right)^{-1}\right)$
$\mathrm{DFS}_{\mathrm{theo,Y_a}} = \frac{1}{K-1}\mathrm{tr}\left(\mathbf{Y}_{\mathrm{a}}\mathbf{Y}_{\mathrm{a}}^{\mathrm{T}}\mathbf{R}^{-1}\right)$
(Identical if the observation operator is linear)
(Estimate)
Theoretical DFS
$\mathrm{DFS}_{\mathrm{theo}} = \mathrm{tr}(\mathbf{HK})$

Figure 1. Overview of equations and their relationships for different degrees of freedom for signal approaches. Samples of vectors $\hat{\mathbf{v}}$, $\hat{\mathbf{r}}$, $\hat{\mathbf{d}}$ and $\mathbf{w}$ are collected from different assimilation times and/or regions where background error statistics are assumed to be sufficiently similar. For details on their equivalence, see the main text.

the resulting DFS estimates depend on the magnitude of the linearization error. Nonlinear observation operators introduce errors into the innovation-based approaches ($\mathrm{DFS}_{\mathrm{act,d}}$, $\mathrm{DFS}_{\mathrm{theo,d}}$ and $\mathrm{DFS}_{\mathrm{theo,d,alt}}$), as discussed in Fowler et al. (2020). Although Fowler et al.'s discussion was presented in the context of variational DA, it also applies to ensemble Kalman filters.

The relationships among the DFS approaches discussed above are summarized in Figure 1. When applying these approaches, it is important to note that—except for the $\mathrm{DFS}_{\mathrm{act,w}}$ approach, which is specifically designed for ETKF systems—all other approaches are applicable to both deterministic and stochastic ensemble Kalman filters. For the innovation-based approaches ($\mathrm{DFS}_{\mathrm{act,d}}$, $\mathrm{DFS}_{\mathrm{theo,d}}$ and $\mathrm{DFS}_{\mathrm{theo,d,alt}}$), there is no difference in their implementation between deterministic and stochastic filters. For the $\mathrm{DFS}_{\mathrm{theo,Y_a}}$ approach, the key distinction lies in how the matrix $\mathbf{Y}_{\mathrm{a}}$ is calculated (see Section 4.2).

In the next section, we discuss how to estimate the DFS in the presence of domain localization, which may introduce differences in the results produced by different DFS approaches (see Figure 7 for an example).

5. Estimating DFS in the Presence of Domain Localization

Domain localization is a commonly used practical technique in ensemble Kalman filters, but it complicates the estimation of the DFS. To address this issue, Hotta and Ota (2021) proposed a strategy (hereafter referred to as the H&O strategy) that constructs a global Kalman gain $(\mathbf{K}_{\mathrm{f}})$ from locally computed Kalman gains and estimates the DFS for each observation as the corresponding diagonal element of the matrix $\mathbf{HK}_{\mathrm{f}}$.

In this work, we propose a novel strategy for estimating the DFS in the presence of domain localization, which is applicable to all DFS approaches considered in this study. Our strategy allows estimation to be performed locally on smaller matrices or vectors, which reduces computational costs and simplifies implementation in operational environments. In general, DFS estimates produced by our strategy are comparable, though not identical, to those obtained using the H&O strategy (see Section 6.4). The theoretical differences between the two strategies are discussed in detail in Appendix F.

Since each local analysis process can be considered an independent analysis process, our idea is to apply the DFS approaches described in Section 4 to each local process and then average the DFS for the same observation over local processes. The steps are described as follows.

1. For each local analysis process, estimate the DFS corresponding to each observation used in that process. We use the notation $\mathrm{DFS}_{i,l}$ to denote the DFS for the ith observation in the lth local analysis process. If the ith

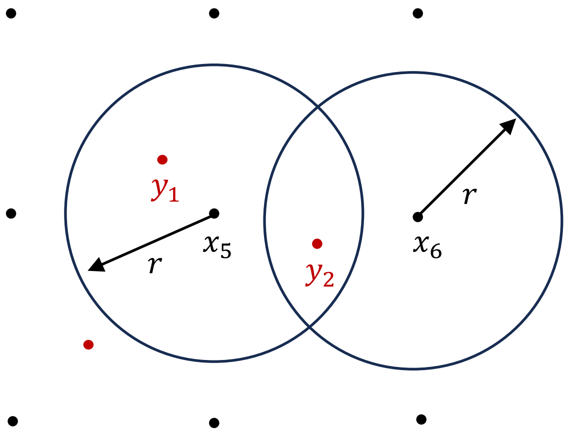

Figure 2. Illustration of the estimation of the influence of each observation with domain localization. The red dots represent the location of observations, and the black dots represent the model grid points. The analyses at x_5 and x_6 are calculated by two independent local data assimilation (DA) processes, each using observations within the localization radius (r). To obtain the influence of the observation y_2 in the entire DA system, we first estimate its influence on x_5 and x_6, respectively, and then average the two estimates.

observation is used in the lth local analysis process, then $\text{DFS}_{i,l}$ may be calculated using any of Equations 15, 17, 19, 25, 27, and 29. Otherwise, $\text{DFS}_{i,l} = 0$.

2. Average the DFS for the same observation over the local analysis processes which use that observation, namely, the DFS for the ith observation in the entire DA system, is computed as $\text{DFS}_i = \frac{1}{n_i}\sum_l^{n_i}\text{DFS}_{i,l}$, where n_i is the number of local analysis processes that use the ith observation.

An illustration of the above steps is given in Figure 2.

To obtain the total influence of all observations, we add up the influence of each observation obtained in Step 2, i.e., $\sum_i^m \text{DFS}_i$. In practice, domain localization is often combined with R-localization, whereby observations farther from the state variable being updated are assigned larger error variances to reduce their influence on that variable (e.g., Hotta & Ota, 2021). Our strategy is applicable with R-localization, but caution is needed when interpreting the results. Under R-localization, an observation may have a large DFS value in a local domain centered near its location, but very small DFS values in domains centered farther away. Averaging these local DFS values, as in Step 2, can therefore lead to a small final DFS value. This might be wrongly interpreted as meaning that the observation is not particularly useful. However, downweighting of observations is a deliberate choice by the user. Moreover, DFS is a relative measure: it is more useful to focus on which observations have larger DFS values than to focus on the exact size of the DFS value, since this varies across DA systems. To provide a more complete assessment, it can be helpful to consider not only the average DFS, but also the distribution of DFS values across all domains.

In the first step, the innovation-based approaches (Equations 25, 27, and 29) are applied as post-processing methods, which use elements of the global innovation, residual and observation-space increment vectors that correspond to the observations used in the local analysis process. The weighting-vector-based approach (Equation 15) and analysis ensemble perturbation approach (Equation 19) are applied along with the assimilation process because they use the intermediate products (the vector $\mathbf{w}$ or the matrix $\mathbf{W}$) generated in each local analysis process.

In addition to domain localization, several other practical techniques are commonly used in ensemble Kalman filters. When covariance localization is applied (Hamill et al., 2001; Houtekamer & Mitchell, 2001), the localized background or observation error covariance matrices should be used in any DFS approach that involves these two matrices. This includes the ensemble perturbation approaches ($\text{DFS}_{\text{theo},\text{Y}_\text{a}}$ and $\text{DFS}_{\text{theo},\text{Y}_\text{b}}$) and the innovation-based approaches ($\text{DFS}_{\text{act,d}}$, $\text{DFS}_{\text{theo,d}}$ and $\text{DFS}_{\text{theo,d,alt}}$). Covariance localization does not affect the $\text{DFS}_{\text{act,w}}$ approach, as the weighting vector $\mathbf{w}$ is already computed using the localized matrices. The effect of covariance inflation (Anderson & Anderson, 1999; Mitchell & Houtekamer, 2000) on the estimation of the DFS can be easily accounted for, as it typically involves applying a scalar factor to the background or analysis error statistics. Finally, when observations are assimilated serially (Anderson, 2001; Dance, 2004; Whitaker & Hamill, 2002) the DFS must be estimated after all observations have been assimilated (Hu et al., 2025).

6. Idealized Data Assimilation Experiments

In this section, we examine the DFS approaches in idealized DA experiments without cycling. We use the local ensemble transform Kalman filter (LETKF; Hunt et al., 2007) as our DA method. Since we do not perform cycling, a forecast model is not needed. Moreover, to further simplify the experiments, we use a linear observation operator and calculate the analysis *error* vector instead of the analysis *state* vector. This avoids the need for background state and observation vectors. Our experiments only require background and observation error vectors, a background ensemble perturbation matrix and an observation error covariance matrix. In addition to using LETKF, we also conducted DA experiments with the ensemble Kalman filter (EnKF; Evensen, 2003) and obtained consistent DFS estimates. Therefore, we present only the LETKF experiments in this study.

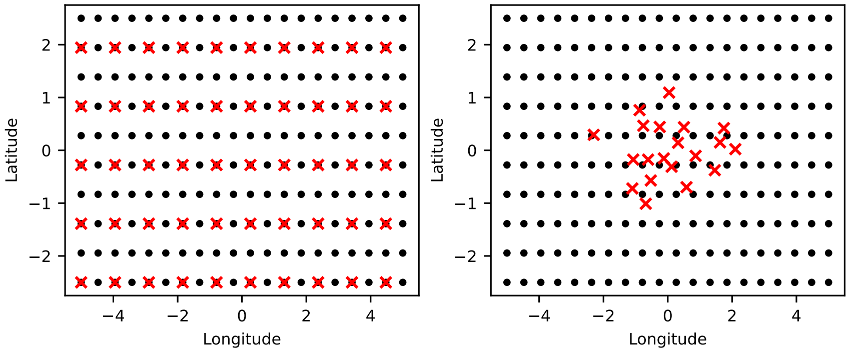

Figure 3. Illustration of two spatial distributions of observations. (Left panel) Observations (red crosses) are regularly distributed on the model grid points (black dots). (Right panel) The location of the observations is randomly selected from a Gaussian distribution.

Since the $\mathrm{DFS}_{\mathrm{act,d}}$ and $\mathrm{DFS}_{\mathrm{act,w}}$ approaches are sensitive to the sample of vectors used in the estimation, we first investigate an appropriate sample size for these approaches and then use it in the subsequent experiments. Then, we investigate the sensitivity of the DFS estimates to ensemble size. After that, we examine our strategy for implementing the DFS approaches in the presence of domain localization. Finally, we look into the spatial variation of the DFS for individual observations due to flow-dependent background error statistics and their relative locations to other observations.

6.1. Experimental Design

Our 20 × 10 model grid points are regularly distributed on a latitude-longitude grid over a region from 2.5°S to 2.5°N and 5°W to 5°E. We consider a region near the equator due to the small differences in grid lengths; the grid length is about 58.5 km in the east-west direction and 61.8 km in the north-south direction. In addition, a small number of model grid points is used to ensure computational efficiency.

We consider two spatial distributions of observations, as shown in Figure 3. In the first case, 50 observations are regularly placed at alternate model grid points, similar to geostationary satellite observations. In the second case, observations are irregularly distributed at locations with longitudes (latitudes) randomly chosen from a Gaussian distribution: the mean is the longitude (latitude) of the grid at the center of the domain, and the standard deviation is 10% of the difference between the maximum and minimum longitudes (latitudes). This distribution is similar to that of radar observations. We further thin the observations to ensure that each grid box contains no more than one observation. After thinning, only 20 of the original 50 observations are left. The Gaussian distribution of observations is used exclusively in Section 6.5, and all the other experiments use the regular distribution.

The true and assumed observation error covariance matrices are identical (i.e., $\mathbf{R} = \mathbf{R}_{\mathrm{t}}$), and both are identity matrices. Nevertheless, it is important to note that the DFS approaches are also applicable when observation errors are correlated. In such cases, the correlation lengthscale affects the influence observations can have on the analysis (e.g., Fowler et al., 2018). The true background error covariance matrix is modeled by the second-order autoregressive (SOAR) correlation function (Daley, 1994; Tabeart et al., 2018), with the (i,j)th element given by

$$\mathbf{B}_{\mathrm{t}}(i,j) = \left(1 + \frac{|\Delta_{i,j}|}{l_{\mathrm{b}}}\right) \exp\left(\frac{-|\Delta_{i,j}|}{l_{\mathrm{b}}}\right),$$

where $l_{\mathrm{b}} = 80$ km is the background error correlation lengthscale and $\Delta_{i,j}$ the great circle distance between two background state variables. The condition number of the resultant matrix $\mathbf{B}_{\mathrm{t}}$ is approximately 1,348.

The background ensemble perturbation matrix $(\mathbf{X}_{\mathrm{b}})$ is generated as follows: each column of the matrix $\mathbf{X}_{\mathrm{b}}$ is given by random values drawn from a Gaussian distribution with mean zero and covariance $\mathbf{B}_{\mathrm{t}}$ (i.e., $\mathbf{X}_{\mathrm{b}}[*,k] \sim \mathcal{N}(\mathbf{0},\mathbf{B}_{\mathrm{t}})$). As the ensemble size (i.e., number of columns of the matrix $\mathbf{X}_{\mathrm{b}}$) increases, the ensemble background error covariance matrix (Equation 1) becomes closer to the true background error covariance matrix,

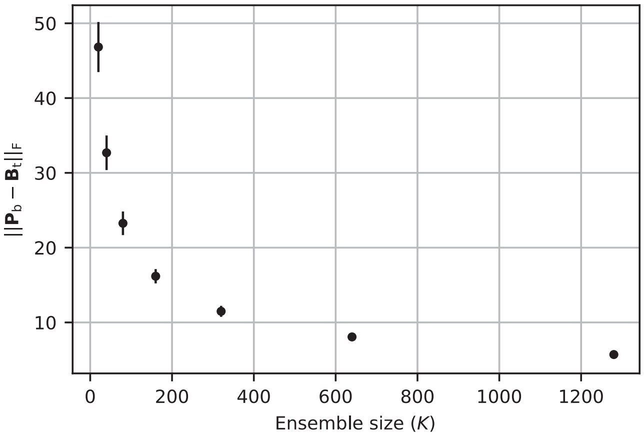

Figure 4. The Frobenius norm of the difference between the ensemble and true background error covariance matrices ($\mathbf{P}_b$ and $\mathbf{B}_t$) as a function of ensemble size (K). For each ensemble size, the mean of 100 estimates of $||\mathbf{P}_b - \mathbf{B}_t||_F$, obtained from different realizations of the matrix $\mathbf{P}_b$, is plotted. The error bar represents the standard deviation of those estimates.

as shown in Figure 4. Due to ensemble sampling error, the sum of the columns of the matrix $\mathbf{X}_b$ may not be a zero vector, resulting in an invalid ensemble perturbation matrix (Livings et al., 2008). To solve this, we remove the bias by subtracting the sample mean from each column.

For observations located on the model grid (left panel of Figure 3), the background ensemble perturbation matrix in observation space ($\mathbf{Y}_b$) is formed by extracting the rows of the matrix $\mathbf{X}_b$ that correspond to the observation locations. For irregularly distributed observations (right panel of Figure 3), the matrix $\mathbf{Y}_b$ is obtained by linearly interpolating the background ensemble perturbations to the observation locations. The interpolation is performed using *scipy.interpolate.LinearNDInterpolator* (Virtanen et al., 2020).

Since our observation operator is linear, calculating the analysis error vector ($\boldsymbol{\varepsilon}_a$) is equivalent to calculating the analysis state vector. The analysis error at each model grid point is computed independently, following Hunt et al. (2007). The details of our experiment are described below.

1. Calculate the global mean innovation vector (defined by Equation 2) as

$$\bar{\mathbf{d}} = \boldsymbol{\varepsilon}_o - \mathbf{H}\boldsymbol{\varepsilon}_b, \tag{32}$$

where $\boldsymbol{\varepsilon}_o \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$ and $\boldsymbol{\varepsilon}_b \sim \mathcal{N}(\mathbf{0}, \mathbf{B}_t)$ are observation and background error vectors, respectively.

2. Perform the local analysis process at each grid point:
   (a) For the lth model grid point, select observations within the localization radius r (see Figure 2). If observations are available, form the local matrices, $\mathbf{R}_l$ and $\mathbf{Y}_{b,l}$, and the local innovation vector, $\bar{\mathbf{d}}_l$. If no observations are within the localization radius, set the analysis error to the background error.
   (b) Calculate

$$\mathbf{C}_l = \mathbf{Y}_{b,l}^\top \mathbf{R}_l^{-1},$$

and perform the following eigendecomposition,

$$(K-1)\mathbf{I} + \mathbf{C}_l\mathbf{Y}_{b,l} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top,$$

where $\mathbf{U}$ is a matrix whose columns are eigenvectors and $\boldsymbol{\Lambda}$ is a diagonal matrix whose elements are eigenvalues. The eigenvalue decomposition is performed using *numpy.linalg.eigh* (Virtanen et al., 2020).

   (c) Calculate the weighting vector as

$$\mathbf{w}_l = \mathbf{U}\boldsymbol{\Lambda}^{-1}\mathbf{U}^\top\mathbf{C}_l\bar{\mathbf{d}}_l,$$

and implement the $\text{DFS}_{\text{act,w}}$ approach using the vector $\mathbf{w}_l$, the vector $\bar{\mathbf{d}}_l$, the matrix $\mathbf{Y}_{b,l}$ and the matrix $\mathbf{R}_l$ (Equation 15).

   (d) Calculate the analysis error as

$$\boldsymbol{\varepsilon}_a[l] = \boldsymbol{\varepsilon}_b[l] + \mathbf{X}_b[l,*]\mathbf{w}_l,$$

where $\boldsymbol{\varepsilon}_a[l]$ and $\boldsymbol{\varepsilon}_b[l]$ denote the lth element of the vectors $\boldsymbol{\varepsilon}_a$ and $\boldsymbol{\varepsilon}_b$ respectively, and $\mathbf{X}_b[l,*]$ denotes the lth row of the matrix $\mathbf{X}_b$.

   (e) Calculate the matrix

$$\mathbf{W}_l = \sqrt{K-1}\mathbf{U}\boldsymbol{\Lambda}^{-1/2}\mathbf{U}^\top,$$

which is used to calculate the matrix $\mathbf{Y}_a$ (see Equation 20) and then to implement the $\text{DFS}_{\text{theo},\mathbf{Y}_a}$ approach.

Table 1
Summary of Experiment Configurations in Each Section, Including Observation Distribution, Vector Sample Size (N), Ensemble Size (K), and Localization Radius (r)

Section	Obs. dist.	N	K	r (km)
6.2	Regular	50, 100, 200, 400, 800	$\mathbf{B}_t$	1,300
6.3	Regular	800	20, 40, 60, 80, 100	1,300
6.4	Regular	800	1,000	50, 100, 200, 400, 800, 1,300
6.5	Regular, Gaussian	800	20, 1,000	1,300

Note. The symbol $\mathbf{B}_t$ indicates that the true background error covariance matrix is used.

3. Calculate the global residual and observation-space increment vectors as

$$\mathbf{r} = \boldsymbol{\varepsilon}_o - \mathbf{H}\boldsymbol{\varepsilon}_a$$

and

$$\mathbf{v} = \overline{\mathbf{d}} - \mathbf{r},$$

whose subvectors are used by the innovation-based approaches ($\text{DFS}_{\text{act,d}}$, $\text{DFS}_{\text{theo,d}}$ and $\text{DFS}_{\text{theo,d,alt}}$).

To obtain samples of the three observation-space vectors ($\overline{\mathbf{d}}$, $\mathbf{r}$ and $\mathbf{v}$) as well as the weighting vector ($\mathbf{w}$), we repeat Steps 1 through 3 a total of N times using different innovation vectors $\overline{\mathbf{d}}$, each calculated using a different sample of observation and background errors (Equation 32).

Table 1 lists the observation distribution, vector sample size (N), ensemble size (K) and localization radius (r) used in the experiments in the following subsections. Note that Section 6.2 uses the true background error covariance matrix ($\mathbf{B}_t$), while the other sections use the ensemble background error covariance matrix. In addition, a localization radius of $r = 1300$ km is equivalent to not having domain localization, as it is longer than the largest separation between any two model grid points.

6.2. Uncertainty in Estimates of Actual DFS

In this section we investigate the role of innovation sampling error in estimating the DFS. As we have seen in Section 3, the definition of the actual DFS (DFS_{act}; Equation 11) involves the true innovation covariance matrix $\mathbb{E}\left[\overline{\mathbf{d}}\,\overline{\mathbf{d}}^{\top}\right]$, which is approximated in practice by a sample mean. In particular, the $\text{DFS}_{\text{act,w}}$ and $\text{DFS}_{\text{act,d}}$ approaches (Equations 14 and 24) rely on finite samples of vectors that are functions of the innovation vector, and are therefore subject to innovation sampling error. Since the ratio of the vector sample size (N) to the number of observations (m) is a more informative quantity than sample size alone (Ledoit & Wolf, 2004), we investigate how the innovation sampling error varies with the ratio N/m. The number of observations is fixed at $m = 50$, as shown in the left panel of Figure 3. We use the true background error covariance matrix ($\mathbf{B}_t$), avoiding additional uncertainty that would arise from using the ensemble background covariance matrix ($\mathbf{P}_b$) that contains ensemble sampling error. The experimental design is summarized in Table 1.

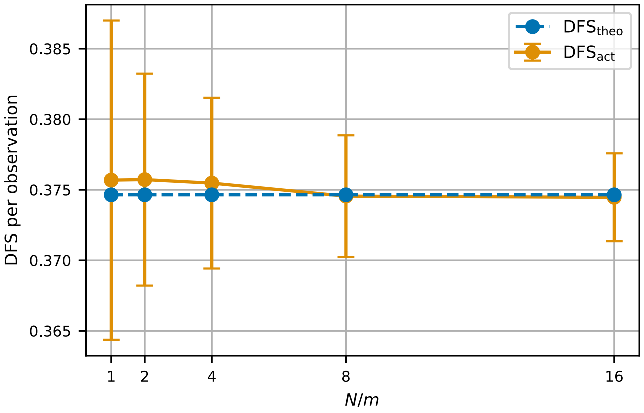

Figure 5. Uncertainty in the estimates of the actual degrees of freedom for signal (DFS) (DFS_{act}) versus the ratio of the vector sample size (N) to the number of observations (m). The theoretical DFS (DFS_{theo}), estimated by the $\text{DFS}_{\text{theo,d}}$ approach, is plotted as a reference and is independent of N/m. DFS_{act} is estimated by the $\text{DFS}_{\text{act,d}}$ approach, and the uncertainty is quantified by the standard deviation of 100 estimates based on different samples of background and observation error vectors. Observations are regularly distributed, as shown in the left panel of Figure 3.

Figure 5 shows the uncertainty in the estimates of the actual DFS as a function of the ratio N/m. We use the $\text{DFS}_{\text{act,d}}$ approach for the estimation, but it should be noted that the $\text{DFS}_{\text{act,w}}$ approach is equivalent to the $\text{DFS}_{\text{act,d}}$ approach (see Section 4.4). In addition, the theoretical DFS (DFS_{theo}; Equation 12) is used as a reference as it does not contain innovation sampling error, and in this plot it is estimated by the $\mathrm{DFS}_{\mathrm{theo,d}}$ approach (Equation 26). As expected, we find that the uncertainty of the estimates of the actual DFS decreases as the ratio N/m increases. At the smallest ratio considered ($N/m = 1$), the standard deviation of the actual DFS estimates is about 0.01, which is approximately 3% of the mean estimate. At the largest ratio considered ($N/m = 16$), the standard deviation of the estimates is less than 1% of the mean value. In later experiments, we will use a ratio of $N/m = 16$ for the $\mathrm{DFS}_{\mathrm{act,w}}$ and $\mathrm{DFS}_{\mathrm{act,d}}$ approaches.

In Figure 5, we used $\mathrm{DFS}_{\mathrm{theo}}$ as a reference as it is not affected by innovation sampling error. Although both the $\mathrm{DFS}_{\mathrm{theo,d}}$ and $\mathrm{DFS}_{\mathrm{theo,d,alt}}$ approaches do involve samples of innovation vectors in their computations, the sampling error effectively cancels. We explain this using the $\mathrm{DFS}_{\mathrm{theo,d,alt}}$ approach as an example, but it can be extended to the $\mathrm{DFS}_{\mathrm{theo,d}}$ approach. Using Equation 31, the first matrix used in the $\mathrm{DFS}_{\mathrm{theo,d,alt}}$ approach (Equation 28) can be approximated as

$$\sum_{j=1}^{N} \hat{\mathbf{v}}_j \hat{\mathbf{d}}_j^{\top} \approx \mathbf{R}^{-1/2}\mathbf{H}\mathbf{K}\left(\sum_{j=1}^{N} \overline{\mathbf{d}}_j \overline{\mathbf{d}}_j^{\top}\right)\mathbf{R}^{-1/2},$$

under the assumption that the matrix $\mathbf{K}$ remains constant across different samples of $\overline{\mathbf{d}}$. The second matrix can be rewritten as

$$\left(\sum_{j=1}^{N} \hat{\mathbf{d}}_j \hat{\mathbf{d}}_j^{\top}\right)^{-1} = \mathbf{R}^{1/2}\left(\sum_{j=1}^{N} \overline{\mathbf{d}}_j \overline{\mathbf{d}}_j^{\top}\right)^{-1}\mathbf{R}^{1/2},$$

assuming that the matrix $\mathbf{R}$ is constant across samples of $\overline{\mathbf{d}}$. When these two matrices are multiplied, the sample covariance estimate, $\sum_{j=1}^{N} \overline{\mathbf{d}}_j \overline{\mathbf{d}}_j^{\top}$, and its inverse, $\left(\sum_{j=1}^{N} \overline{\mathbf{d}}_j \overline{\mathbf{d}}_j^{\top}\right)^{-1}$, cancel out, such that noise in the sample estimate does not affect the final result of the $\mathrm{DFS}_{\mathrm{theo,d,alt}}$ approach. Similarly, by using Equations 30 and 31, we can show that the $\mathrm{DFS}_{\mathrm{theo,d}}$ approach also does not contain innovation sampling error.

One potential problem with the application of the $\mathrm{DFS}_{\mathrm{theo,d}}$ and $\mathrm{DFS}_{\mathrm{theo,d,alt}}$ approaches is that the matrix $\sum_{j=1}^{N} \hat{\mathbf{d}}_j \hat{\mathbf{r}}_j^{\top}$ (or $\sum_{j=1}^{N} \overline{\mathbf{d}}_j \overline{\mathbf{d}}_j^{\top}$) can be ill-conditioned if the ratio N/m is too small (If this ratio is smaller than one, then the matrices are rank deficient and not even invertible). In this case, unrealistic DFS estimates (e.g., negative values) may appear due to the large numerical errors in inverting ill-conditioned matrices. In practice, the use of domain localization can greatly reduce the number of observations considered in each local analysis process, thus helping to achieve favorable N/m ratios.

Our results may provide some ideas for choosing an appropriate ratio N/m in practice. In addition, the innovation sampling error is expected to be larger if the true innovation variance is larger (e.g., Hu & Dance, 2024). A practical challenge in using the innovation-based approaches ($\mathrm{DFS}_{\mathrm{act,d}}$, $\mathrm{DFS}_{\mathrm{theo,d}}$ and $\mathrm{DFS}_{\mathrm{theo,d,alt}}$) and the weighting-vector-based approach ($\mathrm{DFS}_{\mathrm{act,w}}$) is the selection of samples that contain consistent background error statistics. In practice, background error statistics are flow-dependent, and it is generally not possible to collect samples of vectors with exactly the same background error statistics. This introduces an additional source of uncertainty in the DFS estimates.

6.3. Sensitivity to Ensemble Size

The previous experiment uses the true background error covariance matrix ($\mathbf{B}_{\mathrm{t}}$), such that the assumed innovation covariance matrix ($\mathbf{D}$) closely approximates the true innovation covariance matrix $\left(\mathbb{E}\left[\overline{\mathbf{d}}\,\overline{\mathbf{d}}^{\top}\right]\right)$. Therefore, the actual DFS and the theoretical DFS are expected to be similar, as shown in Figure 5. In contrast, this experiment uses the ensemble background error covariance matrix ($\mathbf{P}_{\mathrm{b}}$), under which the assumed and true innovation covariance matrices can differ substantially. The magnitude of this discrepancy depends on ensemble size (K).

As shown in Figure 6, both the actual and theoretical DFS increase with ensemble size. As explained by Hotta and Ota (2021), a possible reason is that the DFS is bounded by the rank of the matrix $\mathbf{P}_{\mathrm{b}}$, which in turn is bounded by $K - 1$. Therefore, a smaller K imposes a smaller upper bound on the DFS. While Figure 6 shows the DFS

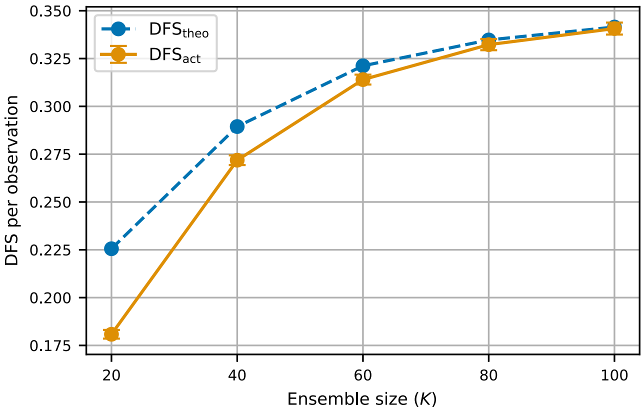

Figure 6. Actual and theoretical degrees of freedom for signal ($\mathrm{DFS_{act}}$ and $\mathrm{DFS_{theo}}$) for the ensemble background error covariance matrix with different ensemble sizes (K). $\mathrm{DFS_{act}}$ is estimated by the $\mathrm{DFS_{act,w}}$ approach, and $\mathrm{DFS_{theo}}$ is estimated by the $\mathrm{DFS_{theo,d,alt}}$ approach, both using a vector sample size of 800. The uncertainty in $\mathrm{DFS_{act,w}}$ is represented by the standard deviation of 100 estimates based on different samples of background and observation error vectors. Observations are regularly distributed, as shown in the left panel of Figure 3.

estimates only up to $K = 100$, we note that as the ensemble size increases to 1,000, the DFS estimates asymptote to those obtained using the true background error covariance matrix. Figure 6 also shows that the difference between the theoretical and actual DFS becomes larger with smaller ensemble sizes. This behavior is expected because the ensemble sampling error in the approximation of the true background error covariance matrix increases as ensemble size decreases. In Figure 6, the actual and theoretical DFS are estimated using the $\mathrm{DFS_{act,w}}$ and $\mathrm{DFS_{theo,d,alt}}$ approaches, respectively. When alternative approaches are used, the results remain consistent.

This experiment may provide us with some insights into the practical use of the actual and theoretical DFS. In this simple experiment, we can use an ensemble of the same order of magnitude in size as the model state variables. However, this is not feasible in practice, and thus, we may not see a good agreement between the actual and theoretical DFS. Instead, we may use the difference between them to inform the quality of a DA system and tune the system parameters (e.g., covariance inflation factor and localization radius). The challenge is that this requires expert knowledge to conjecture whether the difference is mainly due to errors in the observation error covariance matrix, errors in the background error covariance matrix, or errors in the linearization of the observation operator (Fowler et al., 2020). In extreme cases, the background and observation error covariance matrices can be completely wrong while the assumed and true innovation covariance matrices are identical.

6.4. Effect of Domain Localization

Previous experiments did not incorporate domain localization. In this experiment, we use our strategy to implement the DFS approaches under domain localization (see Section 5). The $\mathrm{DFS_{theo,Y_b}}$ approach is not shown, as it produces results identical to those of the $\mathrm{DFS_{theo,Y_a}}$ approach. For comparison, we adopt the H&O strategy, which constructs a global Kalman gain from locally computed gains and then estimates the theoretical DFS using this global gain. In our experiments, the local Kalman gains for the H&O strategy are computed using Equation 5, although they can also be calculated by other ways. The ensemble size and innovation sample size are selected to keep the difference between the actual and theoretical DFS, as well as the innovation sampling error, sufficiently small. This allows us to focus on differences arising from the two different strategies and various DFS approaches.

As shown by Figure 7, the DFS generally decreases as the localization radius increases. Particularly, all DFS approaches produce estimates of approximately 0.5 at a localization radius of 50 km. This radius is shorter than our grid spacing, meaning that each local analysis involves only one grid point and one observation. Given that the background and observation errors have the same standard deviation, and the observation operator is 1 in this scalar case, the Kalman gain and, consequently the DFS should be 0.5. As the localization radius increases further, although more observations are assimilated locally and the total influence of observations is expected to grow, we find that the DFS per observation actually decreases. This occurs because neighboring observations contain similar information, and the information is spread by the background error covariance structure. Beyond a localization radius of 400 km, further increases result in relatively small changes in the DFS per observation. This behavior is expected, as the true background error correlation coefficients approach zero at separation distances of 400 km.

Despite their generally similar behavior, different DFS approaches exhibit some differences. We group them into three panels (panels b, c and d of Figure 7) based on the similarity of their behaviors. In panels (b) and (c), the DFS is around 0.5 at a localization radius of 100 km, for a similar reason as the result at 50 km; both are specific outcomes of our experimental design. As shown in panel (b), when applying our strategy to the $\mathrm{DFS_{theo,d}}$ and $\mathrm{DFS_{act,d}}$ approaches, they produce results consistent with the H&O strategy. This is because both approaches are based on relationships among the innovation, residual, and observation-space increment vectors (Equations 30 and 31), where the Kalman gain can be interpreted as the one constructed by the H&O strategy (see Appendix F6). Panel (c) shows that, under our strategy, the $\mathrm{DFS_{theo,d,alt}}$ approach produces results that differ from those of the $\mathrm{DFS_{theo,d}}$ and $\mathrm{DFS_{act,d}}$ approaches. This is due to the fact that the $\mathrm{DFS_{theo,d,alt}}$ approach does not use the residual

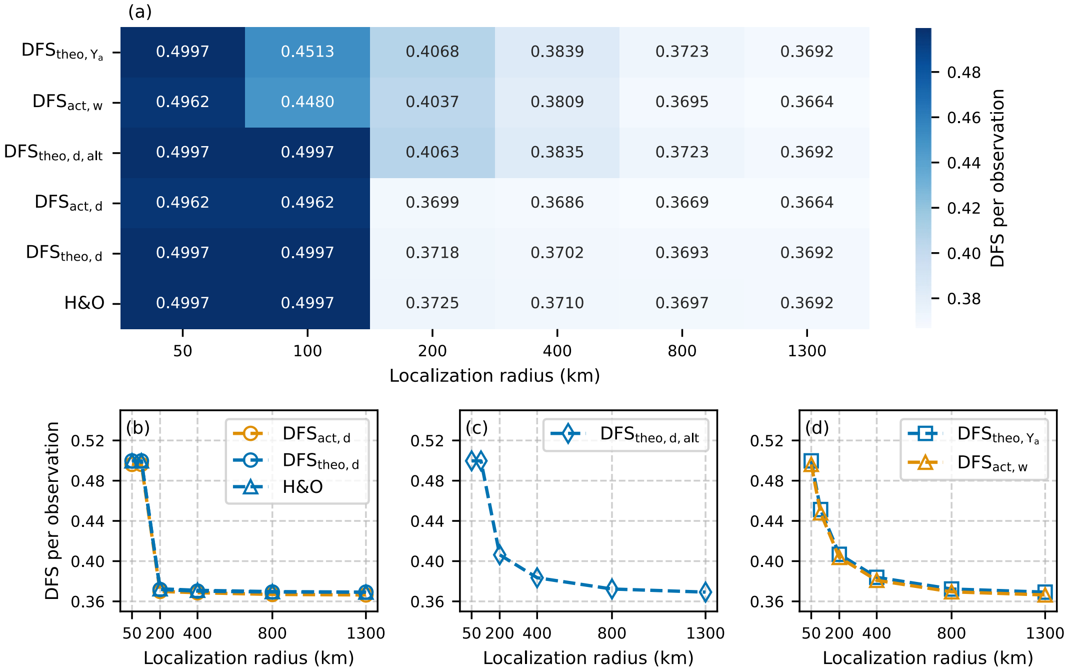

Figure 7. Degrees of freedom for signal (DFS) per observation estimated by various approaches under different localization radii. (a) Heatmap of DFS estimates from each approach. (b) Line plot of DFS estimates from the $\mathrm{DFS}_{\mathrm{theo,d}}$ and $\mathrm{DFS}_{\mathrm{act,d}}$ approaches, as well as the H&O strategy. (c) Line plot of DFS estimates from the $\mathrm{DFS}_{\mathrm{theo,d,alt}}$ approach. (d) Line plot of DFS estimates from the $\mathrm{DFS}_{\mathrm{theo,Y_a}}$ and $\mathrm{DFS}_{\mathrm{act,w}}$ approaches. Observations are regularly distributed, as shown in the left panel of Figure 3.

vector, whereas the other two approaches do (see Appendix F6). Panel (d) shows a difference in the DFS between localization radii of 50 and 100 km. The reason is that, when using our strategy, the $\mathrm{DFS}_{\mathrm{theo,Y_a}}$ and $\mathrm{DFS}_{\mathrm{act,w}}$ approaches use local counterparts of the matrix **HK** (or $\mathbb{E}\left[\bar{\mathbf{d}}\bar{\mathbf{d}}^{\top}\right]\mathbf{D}^{-1}\mathbf{HK}$ for actual DFS), rather than applying the observation operator to a global Kalman gain (see Appendix F). Therefore, our strategy estimates the DFS for each observation by accounting for its various combinations with other observations in each local analysis process. In comparison, the H&O strategy estimates the DFS for each observation by considering that all observations are assimilated in a single global analysis process.

6.5. Spatial Variation in DFS for Individual Observations

This experiment examines how the DFS for an observation varies with background error statistics and the relative locations of other observations. In addition to placing observations directly on the model grid, we also consider randomly selecting observation locations following a Gaussian distribution (Figure 3).

Figure 8 shows the theoretical DFS for each observation (estimated by the $\mathrm{DFS}_{\mathrm{theo,Y_a}}$ approach) in physical space for an ensemble size of 1,000. We find that the spatial mean (shown in the top-right corner of the plot) of the theoretical DFS is larger when observations are regularly distributed than when they are unevenly distributed. This is because the observations in the latter case are more clustered, and therefore, the influence per observation is smaller (Cardinali et al., 2004). We also observe that observations located at the boundary tend to have a slightly greater influence than those located elsewhere. This further highlights that the influence of an observation is affected by the relative locations of others: if there is another observation extremely close to an observation, or if there are many observations surrounding this observation, then this observation is expected to have a smaller influence.

When ensemble size is reduced to 20, the background error statistics become less spatially homogeneous in space due to ensemble sampling error in Equation 1. This means that the background error variance may vary greatly from one grid point to another. In this case, we observe a larger variation of the theoretical DFS for individual

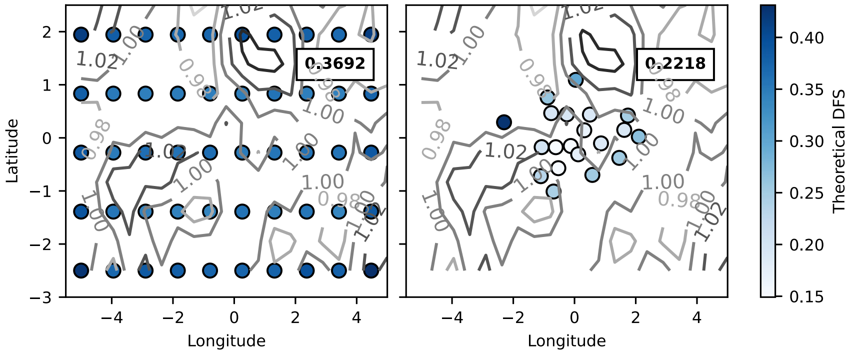

Figure 8. Theoretical degrees of freedom for signal (DFS) for each observation in physical space with two different distributions of observations. The ensemble size is 1,000. Circles represent the locations of the observations, and their color increases with the value of the DFS. Numbers in the upper right corner are spatial averages of the DFS. Contours show the background error standard deviation.

observations (Figure 9). We may explain the variation of the theoretical DFS using Equation 18. Assuming that observation errors are uncorrelated and their variance is the same, the theoretical DFS for the *i*th observation is

$$\mathrm{DFS}_{\mathrm{theo},\mathrm{Y}_\mathrm{a},i} = \frac{1}{\sigma_\mathrm{o}^2(K-1)}\mathbf{Y}_\mathrm{a}[i,*]\cdot\mathbf{Y}_\mathrm{a}[i,*],$$

where σ_o is the observation error standard deviation and $\mathbf{Y}_\mathrm{a}[i,*]$ denotes the row of the matrix $\mathbf{Y}_\mathrm{a}$ corresponding to the *i*th observation. This equation acts as taking the sum of squares of each element of one row of the matrix $\mathbf{Y}_\mathrm{a}$. Using Equation 20, we obtain

$$\mathrm{DFS}_{\mathrm{theo},\mathrm{Y}_\mathrm{a},i} = \frac{1}{\sigma_\mathrm{o}^2(K-1)}\mathbf{Y}_\mathrm{b}[i,*]\mathbf{W}\mathbf{W}^\top\mathbf{Y}_\mathrm{b}^\top[*,i].$$

Using Equation 3, we further obtain

$$\mathrm{DFS}_{\mathrm{theo},\mathrm{Y}_\mathrm{a},i} = \frac{1}{\sigma_\mathrm{o}^2(K-1)}\mathbf{H}[i,*]\mathbf{X}_\mathrm{b}\mathbf{W}\mathbf{W}^\top\mathbf{X}_\mathrm{b}^\top\mathbf{H}^\top[*,i],$$

which shows that the theoretical DFS for the *i*th observation is affected by the background ensemble perturbation of the model state variables used to calculate the model equivalent to that observation. In addition, the effect of the relative locations of observations is reflected by the elements of the matrix **W**.

Compared to the theoretical DFS, the spatial variation of the actual DFS is more complicated to explain. Figure 10 shows the actual DFS for each observation (estimated by the $\mathrm{DFS}_{\mathrm{act,d}}$ approach) in physical space when ensemble

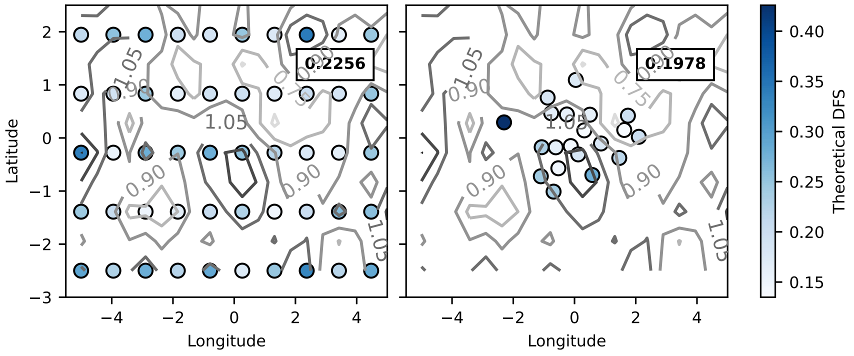

Figure 9. As Figure 8, but for an ensemble size of 20.

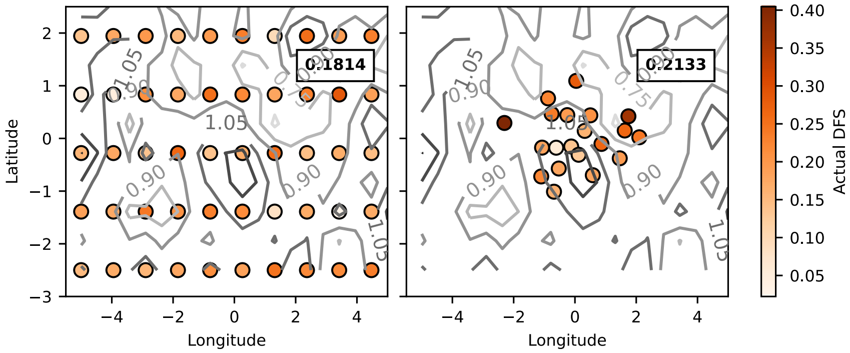

Figure 10. As Figure 8, but for the actual degrees of freedom for signal and an ensemble size of 20.

size is 20. For regularly distributed observations, the spatial mean of the actual DFS is smaller than that of the theoretical DFS (left panel of Figure 9), which is consistent with previous experimental results. However, for observations distributed following a Gaussian distribution, the spatial mean of the actual DFS is found to be larger than that of the theoretical DFS (right panel of Figure 9). Nevertheless, the spatial mean of the actual DFS is still smaller than the "optimal" value obtained with an ensemble size of 1,000 (right panel of Figure 8). The factors influencing the actual DFS estimates are complicated. According to Equation 11, the actual DFS for the *i*th observation is

$$\mathrm{DFS}_{\mathrm{act},i} = \frac{1}{N}\left[\left(\sum_{j}^{N} \bar{\mathbf{d}}_j \bar{\mathbf{d}}_j^{\top}\right)\mathbf{D}^{-1}\mathbf{HK}\right]_{ii},$$

where $[\cdot]_{ii}$ denotes the *i*th diagonal element of the matrix. Using Equations 1 and 6 and assuming uncorrelated observation errors with equal variance, we obtain

$$\mathrm{DFS}_{\mathrm{act},i} = \frac{1}{\sigma_{\mathrm{o}}^{2} N(K-1)}\left[\left(\sum_{j}^{N} \bar{\mathbf{d}}_j \bar{\mathbf{d}}_j^{\top}\right)\mathbf{D}^{-1}\mathbf{H}\mathbf{X}_{\mathrm{a}}\mathbf{X}_{\mathrm{a}}^{\top}\mathbf{H}^{\top}\right]_{ii}.$$

Since $\mathbf{X}_{\mathrm{a}} = \mathbf{X}_{\mathrm{b}}\mathbf{W}$, we have

$$\mathrm{DFS}_{\mathrm{act},i} = \frac{1}{\sigma_{\mathrm{o}}^{2} N(K-1)}\left[\left(\sum_{j}^{N} \bar{\mathbf{d}}_j \bar{\mathbf{d}}_j^{\top}\right)\mathbf{D}^{-1}\mathbf{H}\mathbf{X}_{\mathrm{b}}\mathbf{W}\mathbf{W}^{\top}\mathbf{X}_{\mathrm{b}}^{\top}\mathbf{H}^{\top}\right]_{ii},$$

which is equal to

$$\mathrm{DFS}_{\mathrm{act},i} = \frac{1}{\sigma_{\mathrm{o}}^{2} N(K-1)}\left(\sum_{j}^{N} \bar{\mathbf{d}}_j[i] \cdot \bar{\mathbf{d}}_j^{\top}\right)\mathbf{D}^{-1}\mathbf{H}\mathbf{X}_{\mathrm{b}}\mathbf{W}\mathbf{W}^{\top}\mathbf{X}_{\mathrm{b}}^{\top}\mathbf{H}^{\top}[*, i].$$

This equation indicates that in addition to the factors affecting the theoretical DFS (e.g., relevant background ensemble perturbations and observation locations), the actual DFS for the *i*th observation is also affected by the specific sample of the corresponding innovation.

7. Summary

We investigated how to accurately quantify the influence of observations on the analysis in ensemble-based DA systems using a metric called degrees of freedom for signal (DFS; Equation 10). The DFS measures how much information the analysis has extracted from observations. Existing DFS approaches include the innovation-based approaches (Fowler et al., 2020) and ensemble perturbation approaches (Hotta & Ota, 2021). In addition, we developed a novel weighting-vector-based approach ($\mathrm{DFS}_{\mathrm{act,w}}$) and an alternative formulation of the

Table 2

Comparison of the Advantages of Various Degrees of Freedom for Signal (DFS) Approaches, Including the Weighting-Vector-Based Approach ($DFS_{act,w}$), the Innovation-Based Approaches ($DFS_{act,d}$, $DFS_{theo,d}$ and $DFS_{theo,d,alt}$) and the Ensemble Perturbation Approaches (DFS_{theo,Y_a} and DFS_{theo,Y_b})

Approach	Estimate theoretical DFS	Estimate actual DFS	Applicable to ETKF	Applicable to EnKF	No innovation sampling error
The weighting-vector-based approach		✗	✗		
The innovation-based approaches	✗	✗	✗	✗	
The ensemble perturbation approaches	✗		✗	✗	✗

Note. The symbol ✗ indicates "yes".

innovation-based approach ($\mathrm{DFS}_{\mathrm{theo,d,alt}}$). A summary of the advantages and disadvantages of each DFS approach is provided in Table 2.

Another novel contribution of this work is the development of a general strategy for implementing the DFS approaches in the presence of domain localization. The key idea is that, since local analysis processes are mutually independent, we can apply the DFS approaches separately to each local analysis process to obtain local DFS estimates, and then compute the global DFS for each observation by averaging its local DFS values across the local processes. This novel strategy produces DFS estimates comparable to those produced by the H&O strategy (Hotta & Ota, 2021) that constructs a global Kalman gain for the local analyses. An advantage of our strategy is that it is easier to implement in an operational environment and more computationally efficient, as the DFS estimation is performed locally on small matrices and/or vectors. When R-localization is applied, our strategy may lead to small DFS values, which reflects the user's choice to downweight the observations. The absolute value of the DFS is not particularly meaningful as it varies across different data assimilations systems. We should compare the relatively size of DFS values among observations within the same system. To obtain a more complete assessment, it can also be useful to consider not only the average DFS but also the distribution of DFS values across all local domains.

When quantifying the influence of observations, approaches that use quantities readily available from the DA system are generally preferable due to their computational efficiency (Hu et al., 2025). For example, the innovation-based approaches ($\mathrm{DFS}_{\mathrm{act,d}}$, $\mathrm{DFS}_{\mathrm{theo,d}}$ and $\mathrm{DFS}_{\mathrm{theo,d,alt}}$) use innovation, residual and observation-space increment vectors that are by-products of deterministic or stochastic ensemble Kalman filters (Houtekamer & Zhang, 2016) and can be applied as postprocessing tools. The novel $\mathrm{DFS}_{\mathrm{act,w}}$ approach and the analysis ensemble perturbation approach ($\mathrm{DFS}_{\mathrm{theo,Y_a}}$) instead use intermediate quantities calculated within ETKF systems, and are best incorporated directly into the assimilation workflow. The intermediate quantities can be discarded once the DFS calculation is complete. For example, when applying the $\mathrm{DFS}_{\mathrm{theo,Y_a}}$ approach in the presence of domain localization, we can first compute $\mathbf{Y}_\mathrm{a} = \mathbf{Y}_\mathrm{b}\mathbf{W}$ within each local domain, and then compute the diagonal elements of $1/(K-1)\cdot\mathbf{Y}_\mathrm{a}(\mathbf{Y}_\mathrm{a})^\top\mathbf{R}^{-1}$ as the local DFS contributions for the relevant observations. These contributions are accumulated over local domains for each observation, together with a count of how many times the observation is used. Once this assignment has been done, the matrices $\mathbf{Y}_\mathrm{a}$ and $\mathbf{W}$ can be discarded. In practice, our strategy requires storing two vectors, each with the length equal to the number of observations. In contrast, the H&O strategy requires storing a full Kalman gain matrix. Finally, estimating the actual DFS is generally more computationally expensive than estimating the theoretical DFS, as it requires an estimate of the true innovation covariance matrix based on a sample.

In practical applications, although the observation operator and error statistics are typically fixed for a given observation type, the DFS for each observation may still vary across analysis times and regions due to flow-dependent background error statistics. To deal with this, we may target a weather event of interest and average the DFS for observations over the time period and region of the event. The $\mathrm{DFS}_{\mathrm{act,w}}$ and $\mathrm{DFS}_{\mathrm{act,d}}$ approaches are subject to innovation sampling error, and thus the resulting estimates inherently carry uncertainty in addition to the spatial and temporal variation introduced by the background error statistics. The magnitude of innovation sampling error depends on the ratio of the innovation sample size (N) to the number of observations (m), as well as the magnitude of the elements in the innovation vector (Hu & Dance, 2024). When domain localization is applied, these two approaches operate on local subsets of observations, which improves the N/m

ratio and thereby helps reduce innovation sampling error. The effect of flow-dependent background error statistics on the estimation of DFS warrants further investigation in an operational environment. In addition, the estimation of DFS under highly nonlinear observation operators is another promising direction for future research. Our DFS estimation strategy is being implemented in the Joint Effort for DA Integration (JEDI) framework, and will be used in the Met Office's operational system to explore these future directions.

There are other conceptually different methods that can be used to assess the influence of observations and to identify problems with the assimilation of observations (Diefenbach et al., 2022; Stiller, 2022). Comparing these methods with the DFS approaches would be a valuable direction for future work. In addition, this study focuses solely on the influence of observations on the analysis. However, a greater influence on the analysis does not necessarily correspond to a larger impact on forecast skill, and vice versa. Therefore, the impact of observations on forecast skill should be assessed separately (e.g., Hu et al., 2025).

In summary, the main contributions of this work are the development of novel approaches for estimating the DFS in ensemble-based DA systems and a novel strategy for implementing these approaches in the presence of domain localization. These contributions help to accurately and efficiently quantify the influence of observations, thereby providing valuable guidance for improving both observation networks and DA systems.

Appendix A: True Innovation Covariance

This appendix shows the derivation of the approximation to the true innovation covariance matrix, as given in Equation 8, following Desroziers, Berre, et al. (2005). Let $\mathbf{x}_\mathrm{t} \in \mathbb{R}^n$ be the unknown true model state vector, then the observation vector, the ensemble mean of the background state vector and kth ensemble member of the background state vector can be expressed as

$$\mathbf{y} = h(\mathbf{x}_\mathrm{t}) + \boldsymbol{\varepsilon}_\mathrm{o},$$

$$\overline{\mathbf{x}}_\mathrm{b} = \mathbf{x}_\mathrm{t} + \boldsymbol{\varepsilon}_\mathrm{b} \tag{A1}$$

and

$$\mathbf{x}_{\mathrm{b},k} = \overline{\mathbf{x}}_\mathrm{b} + \mathbf{X}_\mathrm{b}[*,k], \tag{A2}$$

respectively, where $\mathbf{X}_\mathrm{b}[*,k]$ denotes the kth column of the matrix $\mathbf{X}_\mathrm{b}$. Applying the observation operator to Equation A2 and using a first-order Taylor expansion successively around the ensemble mean and true state, we obtain

$$h(\mathbf{x}_{\mathrm{b},k}) \approx h(\mathbf{x}_\mathrm{t}) + \mathbf{H}\boldsymbol{\varepsilon}_\mathrm{b} + \mathbf{H}\mathbf{X}_\mathrm{b}[*,k].$$

Since the matrix $\mathbf{X}_\mathrm{b}$ has zero column sum by definition, taking the ensemble mean gives

$$\overline{h(\mathbf{x}_\mathrm{b})} \approx h(\mathbf{x}_\mathrm{t}) + \mathbf{H}\boldsymbol{\varepsilon}_\mathrm{b}.$$

Using this equation, the ensemble mean of the innovation vector (Equation 2) is approximately

$$\overline{\mathbf{d}} \approx \boldsymbol{\varepsilon}_\mathrm{o} - \mathbf{H}\boldsymbol{\varepsilon}_\mathrm{b}.$$

Then, the true innovation covariance matrix (Equation 9) is approximately

$$\mathbb{E}\left[\overline{\mathbf{d}}\,\overline{\mathbf{d}}^\top\right] \approx \mathbb{E}\left[\boldsymbol{\varepsilon}_\mathrm{o}\boldsymbol{\varepsilon}_\mathrm{o}^\top\right] + \mathbf{H}\mathbb{E}\left[\boldsymbol{\varepsilon}_\mathrm{b}\boldsymbol{\varepsilon}_\mathrm{b}^\top\right]\mathbf{H}^\top$$

where we have assumed that the observation and background errors are mutually uncorrelated, that is, $\mathbb{E}\left[\boldsymbol{\varepsilon}_\mathrm{o}\boldsymbol{\varepsilon}_\mathrm{b}^\top\right] = \mathbb{E}\left[\boldsymbol{\varepsilon}_\mathrm{b}\boldsymbol{\varepsilon}_\mathrm{o}^\top\right] = 0$. Furthermore, since $\mathbb{E}\left[\boldsymbol{\varepsilon}_\mathrm{o}\boldsymbol{\varepsilon}_\mathrm{o}^\top\right] = \mathbf{R}_\mathrm{t}$ is the true observation error covariance matrix and $\mathbb{E}\left[\boldsymbol{\varepsilon}_\mathrm{b}\boldsymbol{\varepsilon}_\mathrm{b}^\top\right] = \mathbf{B}_\mathrm{t}$ is the true background error covariance matrix, we obtain Equation 8 as required.

Appendix B: Derivation of Actual DFS

This appendix shows how to derive the actual DFS (Equation 11) in ensemble Kalman filters. By sequentially substituting Equations 5 and 7 into Equation 10 we obtain

$$\mathrm{DFS} = \mathbb{E}\left[\bar{\mathbf{d}}^{\top}\mathbf{D}^{-1}\mathbf{H}\mathbf{P}_{\mathrm{b}}\mathbf{P}_{\mathrm{b}}^{+}\mathbf{P}_{\mathrm{b}}\mathbf{H}^{\top}\mathbf{D}^{-1}\bar{\mathbf{d}}\right].$$

Since the pseudoinverse (Chapter 5 of Golub & Van Loan, 1996) satisfies

$$\mathbf{P}_{\mathrm{b}}\mathbf{P}_{\mathrm{b}}^{+}\mathbf{P}_{\mathrm{b}} = \mathbf{P}_{\mathrm{b}},$$

we obtain

$$\mathrm{DFS} = \mathbb{E}\left[\bar{\mathbf{d}}^{\top}\mathbf{D}^{-1}\mathbf{H}\mathbf{K}\bar{\mathbf{d}}\right].$$

Since the trace of a scalar is the scalar itself and the trace is invariant under cyclic permutations (Section 2.2 of Bernstein, 2009), we further have

$$\mathrm{DFS} = \mathbb{E}\left[\mathrm{tr}\left(\bar{\mathbf{d}}\,\bar{\mathbf{d}}^{\top}\mathbf{D}^{-1}\mathbf{H}\mathbf{K}\right)\right].$$

Now, assuming that the observation and background error statistics are constant, the above equation becomes

$$\mathrm{DFS} = \mathrm{tr}\left(\mathbb{E}\left[\bar{\mathbf{d}}\,\bar{\mathbf{d}}^{\top}\right]\mathbf{D}^{-1}\mathbf{H}\mathbf{K}\right),$$

as Equation 11.

Appendix C: Derivation of Weighting-Vector-Based Approach

Our new weighting-vector-based approach ($\mathrm{DFS}_{\mathrm{act,w}}$; Equation 14) estimates the actual DFS defined by Equation 11. The derivation proceeds as follows. We first show that the weighting vector calculated in the LETKF system can be rewritten as

$$\mathbf{w} = \frac{1}{K-1}\mathbf{Y}_{\mathrm{b}}^{\top}\widetilde{\mathbf{D}}^{-1}\bar{\mathbf{d}}, \tag{C1}$$

with

$$\widetilde{\mathbf{D}}^{-1} = \left(\frac{1}{K-1}\mathbf{Y}_{\mathrm{b}}\mathbf{Y}_{\mathrm{b}}^{\top} + \mathbf{R}\right)^{-1}$$

being the inverse of the innovation covariance matrix computed using a nonlinear observation operator. Using the Sherman-Morrison-Woodbury formula (Equation 2.1.4 of Golub & Van Loan, 1996), we have

$$\left(\mathbf{R} + \frac{1}{K-1}\mathbf{Y}_{\mathrm{b}}\mathbf{Y}_{\mathrm{b}}^{\top}\right)^{-1} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Y}_{\mathrm{b}}\left[(K-1)\mathbf{I} + \mathbf{Y}_{\mathrm{b}}^{\top}\mathbf{R}^{-1}\mathbf{Y}_{\mathrm{b}}\right]^{-1}\mathbf{Y}_{\mathrm{b}}^{\top}\mathbf{R}^{-1},$$

where we assume that the matrices $\mathbf{R}$ and $\left[(K-1)\mathbf{I} + \mathbf{Y}_{\mathrm{b}}^{\top}\mathbf{R}^{-1}\mathbf{Y}_{\mathrm{b}}\right]$ are nonsingular. Substituting this equation into Equation C1, we obtain

$$\mathbf{w} = \widetilde{\mathbf{P}}_{\mathrm{a}}\mathbf{Y}_{\mathrm{b}}^{\top}\mathbf{R}^{-1}\bar{\mathbf{d}}, \tag{C2}$$

with

$$\widetilde{\mathbf{P}}_{\mathrm{a}} = \left[(K-1)\mathbf{I} + \mathbf{Y}_{\mathrm{b}}^{\top}\mathbf{R}^{-1}\mathbf{Y}_{\mathrm{b}}\right]^{-1}, \tag{C3}$$

which are exactly the equations used in the LETKF (Hunt et al., 2007). Thus, we have shown that the weighting vector can be expressed in terms of Equation C1, which is further used to prove the $\mathrm{DFS}_{\mathrm{act,w}}$ approach.

Using Equation C1, we may write the $\mathrm{DFS}_{\mathrm{act,w}}$ approach as

$$(K-1)\cdot\mathbb{E}\left[\mathbf{w}^{\top}\mathbf{w}\right] = \frac{1}{K-1}\mathbb{E}\left[\overline{\mathbf{d}}^{\top}\widetilde{\mathbf{D}}^{-1}\mathbf{Y}_{\mathrm{b}}\,\mathbf{Y}_{\mathrm{b}}^{\top}\widetilde{\mathbf{D}}^{-1}\overline{\mathbf{d}}\right].$$

Using the properties of the trace and assuming constant observation and background error statistics as in Appendix B, we obtain

$$(K-1)\cdot\mathbb{E}\left[\mathbf{w}^{\top}\mathbf{w}\right] = \frac{1}{K-1}\mathrm{tr}\left(\mathbb{E}\left[\overline{\mathbf{d}}\,\overline{\mathbf{d}}^{\top}\right]\widetilde{\mathbf{D}}^{-1}\mathbf{Y}_{\mathrm{b}}\,\mathbf{Y}_{\mathrm{b}}^{\top}\widetilde{\mathbf{D}}^{-1}\right).$$

Assuming a linear observation operator, we finally obtain

$$(K-1)\cdot\mathbb{E}\left[\mathbf{w}^{\top}\mathbf{w}\right] = \mathrm{tr}\left(\mathbb{E}\left[\overline{\mathbf{d}}\,\overline{\mathbf{d}}^{\top}\right]\mathbf{D}^{-1}\mathbf{H}\mathbf{K}\right),$$

which shows that the $\mathrm{DFS}_{\mathrm{act,w}}$ approach estimates the actual DFS.

The equation used to estimate the DFS for a subset of observations ($\mathrm{DFS}_{\mathrm{act,w},i}$, Equation 15) can be proved in a similar way. Using Equation C1, we obtain

$$\mathbb{E}\left[\left(\overline{\mathbf{d}} - \mathbf{Y}_{\mathrm{b}}\mathbf{w}\right)^{\top}\mathbf{R}^{-1}\mathbf{Y}_{\mathrm{b}}\mathbf{w}\right] = \mathbb{E}\left[\overline{\mathbf{d}}^{\top}\mathbf{D}^{-1}\mathbf{H}\mathbf{K}\overline{\mathbf{d}}\right]$$

for linear observation operators. Again, using the properties of the trace and assuming constant observation and background error statistics, we obtain

$$\mathbb{E}\left[\left(\overline{\mathbf{d}} - \mathbf{Y}_{\mathrm{b}}\mathbf{w}\right)^{\top}\mathbf{R}^{-1}\mathbf{Y}_{\mathrm{b}}\mathbf{w}\right] = \mathrm{tr}\left(\mathbb{E}\left[\overline{\mathbf{d}}\,\overline{\mathbf{d}}^{\top}\right]\mathbf{D}^{-1}\mathbf{H}\mathbf{K}\right),$$

which is the actual DFS. Since the vectors $\overline{\mathbf{d}} - \mathbf{Y}_{\mathrm{b}}\mathbf{w}$ and $\mathbf{Y}_{\mathrm{b}}\mathbf{w}$ are in observation space, we can isolate the DFS contribution of a subset of observations as

$$\begin{aligned}\mathrm{DFS}_{\mathrm{act,w},i} &= \frac{1}{N}\sum_{j=1}^{N}\left(\mathbf{\Pi}_i\overline{\mathbf{d}}_j - \mathbf{\Pi}_i\,\mathbf{Y}_{\mathrm{b},j}\mathbf{w}_j\right)^{\top}\mathbf{\Pi}_i\mathbf{R}^{-1}\mathbf{\Pi}_i{}^{\top}\mathbf{\Pi}_i\,\mathbf{Y}_{\mathrm{b},j}\mathbf{w}_j\\ &= \frac{1}{N}\sum_{j=1}^{N}\left(\overline{\mathbf{d}}_j - \mathbf{Y}_{\mathrm{b},j}\mathbf{w}_j\right)^{\top}\mathbf{S}_i\mathbf{R}^{-1}\mathbf{S}_i\,\mathbf{Y}_{\mathrm{b},j}\mathbf{w}_j,\end{aligned}$$

where we have $\mathbf{S}_i = \mathbf{\Pi}_i^{\top}\mathbf{\Pi}_i$ as defined. Since $\mathbf{S}_i$ commutes with $\mathbf{R}^{-1}$ (as $\mathbf{R}^{-1}$ is typically block-diagonal) and $\mathbf{S}_i\mathbf{S}_i = \mathbf{S}_i$ (Desroziers, Brousseau, & Chapnik, 2005), we finally recover Equation 15, as required.

Appendix D: Calculation of Innovation, Residual and Observation-Space Increment Vectors

In Equations 21 and 22, we use $h(\overline{\mathbf{x}}_{\mathrm{a}})$ and $h(\overline{\mathbf{x}}_{\mathrm{b}})$ rather than $\overline{h(\mathbf{x}_{\mathrm{a}})}$ and $\overline{h(\mathbf{x}_{\mathrm{b}})}$ to calculate the residual and observation-space increment vectors because $\overline{\mathbf{x}}_{\mathrm{a}}$ and $\overline{\mathbf{x}}_{\mathrm{b}}$ are used in the analysis equation (Equation 7). To obtain the relationship between $h(\bar{\mathbf{x}}_\mathrm{a})$ and $h(\bar{\mathbf{x}}_\mathrm{b})$, we can apply a nonlinear observation operator to both sides of the analysis equation and expand the right-hand side using the Taylor expansion.

If the observation operator is linear, then $\overline{h(\mathbf{x}_\mathrm{a})} = h(\bar{\mathbf{x}}_\mathrm{a})$ and $\overline{h(\mathbf{x}_\mathrm{b})} = h(\bar{\mathbf{x}}_\mathrm{b})$. If the observation operator is nonlinear, we can use the Taylor expansion to obtain the approximations, $\overline{h(\mathbf{x}_\mathrm{a})} \approx h(\bar{\mathbf{x}}_\mathrm{a})$ and $\overline{h(\mathbf{x}_\mathrm{b})} \approx h(\bar{\mathbf{x}}_\mathrm{b})$, where the approximation error depends on the size of the ensemble spread and the degree of nonlinearity of the observation operator.

Depending on the DA system, it is also possible to calculate the innovation, residual and observation-space increment vectors using a selected ensemble member (the control member) instead of the ensemble mean. These two different choices should give us the same value of the theoretical DFS because it is determined by the assumed background and observation error statistics only (Equation 12). However, the estimation of the actual DFS also requires a sample estimate of the true innovation covariance matrix, $\mathbb{E}\left[\bar{\mathbf{d}}\,\bar{\mathbf{d}}^\top\right]$. In ensemble-based DA, the background ensemble mean is defined as the first guess (Evensen, 2003). The control member can be considered as adding a perturbation to the first guess. Consequently, innovation vectors calculated using the ensemble mean and the control member contain different error statistics, resulting in different values of the actual DFS.

Appendix E: Derivation of Alternative Innovation-Based Approach

This appendix presents the derivation of the alternative innovation-based approach ($\mathrm{DFS}_{\mathrm{theo,d,alt}}$; Equation 28), which estimates the theoretical DFS (Equation 12). Using Equations 23 and 31, we have

$$\hat{\mathbf{v}}\hat{\mathbf{d}}^\top \approx \mathbf{R}^{-1/2}\mathbf{H}\mathbf{K}\mathbf{R}^{1/2}\hat{\mathbf{d}}\hat{\mathbf{d}}^\top.$$

Taking the statistical expectation of both sides while assuming the matrices $\mathbf{R}^{-1/2}$, $\mathbf{H}$ and $\mathbf{K}$ are invariant, we obtain

$$\mathbb{E}\left[\hat{\mathbf{v}}\hat{\mathbf{d}}^\top\right]\left(\mathbb{E}\left[\hat{\mathbf{d}}\hat{\mathbf{d}}^\top\right]\right)^{-1} \approx \mathbf{R}^{-1/2}\mathbf{H}\mathbf{K}\mathbf{R}^{1/2},$$

where the matrix $\mathbb{E}\left[\hat{\mathbf{d}}\hat{\mathbf{d}}^\top\right]$ is assumed to be invertible. Using the cyclic property of the trace, we can then show that the $\mathrm{DFS}_{\mathrm{theo,d,alt}}$ approach estimates the theoretical DFS.

Appendix F: Comparison of Two Strategies

In this appendix, we clarify the difference between our strategy (Section 5) and the H&O strategy (Hotta & Ota, 2021) for estimating the DFS in the presence of domain localization. For simplicity, we focus on the estimation of the theoretical DFS. We let $\{x_i | i = 1, \ldots, n\}$ denote n grid points and $\{y_j | j = 1, \ldots, m\}$ denote m observations. Furthermore, we assume that the analyses for each point are calculated independently using $m_{x_i} < m$ local observations.

F1. H&O Strategy

The key idea behind the H&O strategy is to construct the global Kalman gain by aggregating local Kalman gains. Here, these local Kalman gains are calculated using background error statistics, but they can alternatively be derived from analysis error statistics. For the grid point x_i, the jth element of the local Kalman gain is

$$\mathbf{k}_{x_i}[j] = \sum_{k=1}^{K} \mathbf{X}_\mathrm{b}[i,k]\mathbf{S}_{x_i}[k,j], \tag{F1}$$

where $\mathbf{X}_\mathrm{b}[i,k]$ denotes the (i,k)th element of the matrix $\mathbf{X}_\mathrm{b}$, and $\mathbf{S}_{x_i}$ is the $N \times m_{x_i}$ matrix given by

$$\mathbf{S}_{x_i} = (\mathbf{Y}_\mathrm{b}[\mathcal{I}^y_{x_i},*])^\top\left(\mathbf{Y}_\mathrm{b}[\mathcal{I}^y_{x_i},*](\mathbf{Y}_\mathrm{b}[\mathcal{I}^y_{x_i},*])^\top + (K-1)\mathbf{R}[\mathcal{I}^y_{x_i},\mathcal{I}^y_{x_i}]\right)^{-1}, \tag{F2}$$

with $\mathcal{I}^y_{x_i} \in \mathbb{R}^{m_{x_i}}$ being a vector containing the indices of observations used to create the analysis at x_i, $\mathbf{Y}_\mathrm{b}[\mathcal{I}^y_{x_i},*] \in \mathbb{R}^{m_{x_i}\times K}$ being the submatrix of the matrix $\mathbf{Y}_\mathrm{b}$, consisting of the rows selected according to $\mathcal{I}^y_{x_i}$, and $\mathbf{R}[\mathcal{I}^y_{x_i},\mathcal{I}^y_{x_i}] \in \mathbb{R}^{m_{x_i}\times m_{x_i}}$ being the submatrix of the matrix $\mathbf{R}$, consisting of the rows and columns selected according to $\mathcal{I}^y_{x_i}$.

Let $\mathbf{K}_\mathrm{f} \in \mathbb{R}^{n\times m}$ be a global Kalman gain, whose ith row is given by

$$\mathbf{K}_\mathrm{f}[i,\mathcal{I}^y_{x_i}[j']] = \mathbf{k}_{x_i}[j'] \tag{F3}$$

for $j' = 1,\dots,m_{x_i}$, with all other columns set to 0. For example, if we have four equally spaced grid points on a circle with periodic boundary conditions, direct observations and a localization radius of one grid point, then we have $\mathcal{I}^y_{x_1} = \{4,1,2\}$, $\mathcal{I}^y_{x_2} = \{1,2,3\}$, $\mathcal{I}^y_{x_3} = \{2,3,4\}$ and $\mathcal{I}^y_{x_4} = \{3,4,1\}$, and the global Kalman gain is

$$\mathbf{K}_{\mathrm{f},n=4} = \begin{bmatrix} \mathbf{k}_{x_1}[2] & \mathbf{k}_{x_1}[3] & 0 & \mathbf{k}_{x_1}[1] \\ \mathbf{k}_{x_2}[1] & \mathbf{k}_{x_2}[2] & \mathbf{k}_{x_2}[3] & 0 \\ 0 & \mathbf{k}_{x_3}[1] & \mathbf{k}_{x_3}[2] & \mathbf{k}_{x_3}[3] \\ \mathbf{k}_{x_4}[3] & 0 & \mathbf{k}_{x_4}[1] & \mathbf{k}_{x_4}[2] \end{bmatrix}$$

as given by Equation F3.

After obtaining the matrix $\mathbf{K}_\mathrm{f}$, the DFS for observation y_j is the jth diagonal elements of $\mathbf{HK}_\mathrm{f}$, which is

$$\mathrm{DFS}_{y_j} = \sum_{i\in\mathcal{I}^x_{h_j}} \sum_{k=1}^{K} \mathbf{H}[j,i]\mathbf{X}_\mathrm{b}[i,k]\mathbf{S}_{x_i}[k,j'], \tag{F4}$$

where $\mathbf{H} \in \mathbb{R}^{m\times n}$ is the linear (or linearized) observation operator, $\mathcal{I}^x_{h_j}$ is a vector containing the indices of grid points that are used to compute the model equivalence to observation y_j, and the index $j' \in \{1,2,\dots,m_{x_i}\}$ satisfies $\mathcal{I}^y_{x_i}[j'] = j$.

F2. Our Strategy

Our strategy is applicable to all DFS approaches considered in this work. To facilitate comparison with the H&O strategy, we illustrate it here using the $\mathrm{DFS}_{\mathrm{theo},\mathrm{Y_b}}$ approach (Equation 16). The application of our strategy to innovation-based approaches ($\mathrm{DFS}_{\mathrm{theo,d}}$ and $\mathrm{DFS}_{\mathrm{theo,d,alt}}$) is discussed later in Appendix F6.

Our strategy starts with calculating a local matrix as

$$\mathbf{Q}_{x_i}[j',j'] = \sum_{k=1}^{K} \mathbf{Y}_\mathrm{b}[\mathcal{I}^y_{x_i}[j'],k]\,\mathbf{S}_{x_i}[k,j'] \tag{F5}$$

for $j' = 1,\dots,m_{x_i}$. The diagonal elements of the matrix $\mathbf{Q}_{x_i}$ are the local DFS for each observation used in the local analysis process for x_i. Since the same observation may be used in multiple local analysis processes, we average the DFS for the same observation across the local analysis processes which use that observation. Specifically, the total DFS for observation y_j is defined as

$$\mathrm{DFS}^*_{y_j} = \frac{1}{n_j} \sum_{i'\in\mathcal{I}^x_{y_j}} \sum_{k=1}^{K} \mathbf{Y}_\mathrm{b}[j,k]\mathbf{S}_{x_{i'}}[k,j'], \tag{F6}$$

where $\mathcal{I}^{x}_{y_j} \in \mathbb{R}^{n_j}$ is a vector containing the indices of grid points whose analysis is created using observation y_j, and the index $j' \in \{1, 2, \ldots, m_{x_{i'}}\}$ is such that $\mathcal{I}^{y}_{x_{i'}}[j'] = j$.

F3. Linear Observation Operators

The difference between our strategy and the H&O strategy is affected by whether a linear or nonlinear observation operator is used. In the linear case, the matrix $\mathbf{Y}_\mathrm{b}$ can be expressed as

$$\mathbf{Y}_\mathrm{b}[j,k] = \sum_{i \in \mathcal{I}^{x}_{h_j}} \mathbf{H}[j,i]\mathbf{X}_\mathrm{b}[i,k]. \tag{F7}$$

Substituting Equation F7 into Equation F6, the DFS for observation y_j under our strategy becomes

$$\mathrm{DFS}^{*}_{y_j} = \frac{1}{n_j} \sum_{i' \in \mathcal{I}^{x}_{y_j}} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}^{x}_{h_j}} \mathbf{H}[j,i]\mathbf{X}_\mathrm{b}[i,k]\mathbf{S}_{x_{i'}}[k,j'], \tag{F8}$$

where the index $j' \in \{1, 2, \ldots, m_{x_{i'}}\}$ satisfies $\mathcal{I}^{y}_{x_{i'}}[j'] = j$. Comparing Equation F8 with Equation F4, we observe that a key difference lies in the choice of matrices $\mathbf{S}_{x_i}$. The H&O strategy uses $\mathbf{S}_{x_i}$ with $i \in \mathcal{I}^{x}_{h_j}$, that is, the matrices associated with the grid points linked to observation y_j through the observation operator (Equation F4). In contrast, our strategy employs $\mathbf{S}_{x_{i'}}$ with $i' \in \mathcal{I}^{x}_{y_j}$, that is, the matrices corresponding to the grid points located within the localization length from observation y_j (Equation F8).

F4. Nonlinear Observation Operators

The above discussion assumes that the observation operator is linear. However, in many practical applications, the observation operator is nonlinear. In such cases, the relationship between $\mathbf{Y}_\mathrm{b}$ and $\mathbf{X}_\mathrm{b}$ can no longer be expressed in the form of Equation F7, and the difference between our strategy and the H&O strategy may become more pronounced.

When the nonlinear observation operator involves multiple grid points (e.g., for satellite radiances and radio occultation observations), our strategy uses the nonlinear operator directly and the H&O strategy applies a linearized operator to the constructed global Kalman gain. Therefore, the difference between the two strategies depends on the linearization error of the observation operator. If the matrix $\mathbf{Y}_\mathrm{b}$ in Equation F5 is computed using a linearized observation operator (noting that the matrix $\mathbf{S}_{x_i}$ can still be computed using the nonlinear observation operator), the difference between the results of the two strategies can be reduced.

F5. When Do the Two Strategies Produce Identical Results?

Although our strategy and the H&O strategy do not generally produce exactly the same results, they can be shown to be equivalent under certain conditions. For example, if model grid points are observed directly and the observation operator is linear, then Equation F4 simplifies to

$$\mathrm{DFS}_{y_i} = \sum_{k=1}^{K} \mathbf{H}[i,i]\mathbf{X}_\mathrm{b}[i,k]\mathbf{S}_{x_i}[k,j'], \tag{F9}$$

where the index $j' \in \{1, 2, \ldots, m_{x_i}\}$ satisfies $\mathcal{I}^{y}_{x_i}[j'] = i$. Similarly, Equation F6 becomes

$$\mathrm{DFS}^{*}_{y_i} = \frac{1}{n_{y_i}} \sum_{i' \in \mathcal{I}^{x}_{y_i}} \sum_{k=1}^{K} \mathbf{H}[i,i]\mathbf{X}_\mathrm{b}[i,k]\mathbf{S}_{x_{i'}}[k,j'], \tag{F10}$$

where the index $j' \in \{1, 2, \ldots, m_{x_{i'}}\}$ satisfies $\mathcal{I}^{y}_{x_{i'}}[j'] = j$. If all matrices $\mathbf{S}_{x_{i'}}$ with $i' \in \mathcal{I}^{x}_{y_j}$ are equal to $\mathbf{S}_{x_i}$, then Equation F10 reduces to Equation F9.

According to Equation F2, the equivalence of matrices $\mathbf{S}_{x_i}$ requires that both $\mathbf{Y}_{\mathrm{b}}[\mathcal{I}^{y}_{x_i}, *]$ and $\mathbf{R}[\mathcal{I}^{y}_{x_i}, \mathcal{I}^{y}_{x_i}]$ remain consistent across local analysis processes that include observation y_i. This requires several conditions, including spatially heterogeneous background error statistics, identical observation operators, identical observation error statistics, and the absence of observation error inflation.

F6. In the Case of Innovation-Based Approaches

This appendix discusses the application of our strategy to the $\mathrm{DFS}_{\mathrm{theo,d}}$ and $\mathrm{DFS}_{\mathrm{theo,d,alt}}$ approaches. Under our strategy, the $\mathrm{DFS}_{\mathrm{theo,d}}$ approach operates on subvectors of the innovation, residual and increment vectors. In this case, the relationships between the residual and innovation vectors (Equation 30) and between the increment and innovation vectors (Equation 31) become

$$\mathbf{r}[\mathcal{I}^{y}_{x_i}] \approx \overline{\mathbf{d}}[\mathcal{I}^{y}_{x_i}] - \mathbf{H}[\mathcal{I}^{y}_{x_i}, *]\,\mathbf{K}_{\mathrm{f}}\overline{\mathbf{d}} \tag{F11}$$

and

$$\mathbf{v}[\mathcal{I}^{y}_{x_i}] \approx \mathbf{H}[\mathcal{I}^{y}_{x_i}, *]\,\mathbf{K}_{\mathrm{f}}\overline{\mathbf{d}}, \tag{F12}$$

where unnormalized vectors are used for brevity. Substituting Equations F11 and F12 into Equation 26, we find that the local DFS for individual observations corresponds to the diagonal elements of the matrix

$$\mathbf{H}[\mathcal{I}^{y}_{x_i}, *]\,\mathbf{K}_{\mathrm{f}}.$$

The global DFS for each observation is then computed as the average of the corresponding local DFS estimates across all relevant local analysis processes, which is

$$\mathrm{DFS}^{*}_{\mathrm{theo,d}} = \frac{1}{n_j} \sum_{i' \in \mathcal{I}^{x}_{y_j}} \left[\mathbf{H}[\mathcal{I}^{y}_{x_{i'}}, *]\mathbf{K}_{\mathrm{f}}\right]_{j'j'}, \tag{F13}$$

where the index $j' \in \{1, 2, \ldots, m_{x_{i'}}\}$ satisfies $\mathcal{I}_{x_{i'}}{}^{y}[j'] = j$. Equation F13 shows that the $\mathrm{DFS}_{\mathrm{theo,d}}$ approach provides DFS estimates based on the constructed global Kalman gain as the H&O strategy does. Therefore, our strategy produces results that are more aligned with the H&O strategy when using the $\mathrm{DFS}_{\mathrm{theo,d}}$ approach compared to other approaches (see Figure 7).

The $\mathrm{DFS}_{\mathrm{theo,d,alt}}$ approach produces different results than the $\mathrm{DFS}_{\mathrm{theo,d}}$ approach under our strategy. Substituting Equation F12 into Equation 28, we obtain the local matrix for the $\mathrm{DFS}_{\mathrm{theo,d,alt}}$ approach, which is

$$\sum_{j=1}^{N} \mathbf{H}[\mathcal{I}^{y}_{x_i}, *]\,\mathbf{K}_{\mathrm{f}}\overline{\mathbf{d}}_j\overline{\mathbf{d}}[\mathcal{I}^{y}_{x_i}]^{\top} \left(\sum_{j=1}^{N} \overline{\mathbf{d}}[\mathcal{I}^{y}_{x_i}]\overline{\mathbf{d}}[\mathcal{I}^{y}_{x_i}]^{\top} \right)^{-1},$$

where the innovation vector cannot be canceled out. This highlights a key distinction between the $\mathrm{DFS}_{\mathrm{theo,d}}$ and $\mathrm{DFS}_{\mathrm{theo,d,alt}}$ approaches when applied using our strategy.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

The Python code used to perform the data assimilation and degrees of freedom for signal experiments described in this study is openly available at https://github.com/Hu831/DADFS.git and archived on Zenodo at https://doi.org/10.5281/zenodo.15388900 (Hu, 2025). The code is distributed under the MIT License.

Acknowledgments

This work was funded in part by the UK NERC (Natural Environment Research Council) NCEO (National Centre for Earth Observation) and the UK Met Office. We sincerely thank the reviewers for their constructive and insightful suggestions that helped to improve the quality of this work.

References

Anderson, J. L. (2001). An ensemble adjustment Kalman filter for data assimilation. *Monthly Weather Review*, *129*(12), 2884–2903. https://doi.org/10.1175/1520-0493(2001)129⟨2884:AEAKFF⟩2.0.CO;2

Anderson, J. L., & Anderson, S. L. (1999). A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Monthly Weather Review*, *127*(12), 2741–2758. https://doi.org/10.1175/1520-0493(1999)127⟨2741:AMCIOT⟩2.0.CO;2

Bai, Z., Fahey, G., & Golub, G. (1996). Some large-scale matrix computation problems. *Journal of Computational and Applied Mathematics*, *74*(1), 71–89. https://doi.org/10.1016/0377-0427(96)00018-0

Ballard, S. P., Li, Z., Simonin, D., & Caron, J.-F. (2016). Performance of 4D-Var NWP-based nowcasting of precipitation at the Met Office for summer 2012. *Quarterly Journal of the Royal Meteorological Society*, *142*(694), 472–487. https://doi.org/10.1002/qj.2665

Bauer, P., Thorpe, A., & Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, *525*(7567), 47–55. https://doi.org/10.1038/nature14956

Bernstein, D. S. (2009). *Matrix mathematics: Theory, facts, and formulas* (2nd ed.). Princeton University Press. https://doi.org/10.1515/9781400833344

Bishop, C. H., Etherton, B. J., & Majumdar, S. J. (2001). Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects. *Monthly Weather Review*, *129*(3), 420–436. https://doi.org/10.1175/1520-0493(2001)129⟨0420:ASWTET⟩2.0.CO;2

Brousseau, P., Berre, L., Bouttier, F., & Desroziers, G. (2012). Flow-dependent background-error covariances for a convective-scale data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, *138*(663), 310–322. https://doi.org/10.1002/qj.920

Cardinali, C., Pezzulli, S., & Andersson, E. (2004). Influence matrix diagnostic of a data assimilation system (Vol. *450*, p. 23). https://doi.org/10.21957/v87lmti8e

Carrera, M. L., Bélair, S., & Bilodeau, B. (2015). The Canadian land data assimilation System (CaLDAS): Description and synthetic evaluation Study. *Journal of Hydrometeorology*, *16*(3), 1293–1314. https://doi.org/10.1175/JHM-D-14-0089.1

Chapnik, B., Desroziers, G., Rabier, F., & Talagrand, O. (2006). Diagnosis and tuning of observational error in a quasi-operational data assimilation setting. *Quarterly Journal of the Royal Meteorological Society*, *132*(615), 543–565. https://doi.org/10.1256/qj.04.102

Collard, A. D. (2007). Selection of IASI channels for use in numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, *133*(629), 1977–1991. https://doi.org/10.1002/qj.178

Daley, R. (1994). *Atmospheric data analysis*. Cambridge University Press.

Dance, S. (2004). Issues in high resolution limited area data assimilation for quantitative precipitation forecasting. *Physica D: Nonlinear Phenomena*, *196*(1), 1–27. https://doi.org/10.1016/j.physd.2004.05.001

Desroziers, G., Berre, L., Chapnik, B., & Poli, P. (2005). Diagnosis of observation, background and analysis-error statistics in observation space. *Quarterly Journal of the Royal Meteorological Society*, *131*(613), 3385–3396. https://doi.org/10.1256/qj.05.108

Desroziers, G., Brousseau, P., & Chapnik, B. (2005). Use of randomization to diagnose the impact of observations on analyses and forecasts. *Quarterly Journal of the Royal Meteorological Society*, *131*(611), 2821–2837. https://doi.org/10.1256/qj.04.151

Desroziers, G., & Ivanov, S. (2001). Diagnosis and adaptive tuning of observation-error parameters in a variational assimilation. *Quarterly Journal of the Royal Meteorological Society*, *127*(574), 1433–1452. https://doi.org/10.1002/qj.49712757417

Diefenbach, T., Craig, G., Keil, C., Scheck, L., & Weissmann, M. (2022). Partial analysis increments as diagnostic for LETKF data assimilation systems. *Quarterly Journal of the Royal Meteorological Society*, *149*(752), 740–756. https://doi.org/10.1002/qj.4419

Evensen, G. (2003). The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dynamics*, *53*(4), 343–367. https://doi.org/10.1007/s10236-003-0036-9

Fischer, C., Montmerle, T., Berre, L., Auger, L., & Ştefănescu, S. E. (2005). An overview of the variational assimilation in the ALADIN/France numerical weather-prediction system. *Quarterly Journal of the Royal Meteorological Society*, *131*(613), 3477–3492. https://doi.org/10.1256/qj.05.115

Fisher, M. (2003). Estimation of entropy reduction and degrees of freedom for signal for large variational analysis systems (Vol. *397*, p. 18). https://doi.org/10.21957/2bec9m38o

Fowler, A. M., Dance, S. L., & Waller, J. A. (2018). On the interaction of observation and prior error correlations in data assimilation. *Quarterly Journal of the Royal Meteorological Society*, *144*(710), 48–62. https://doi.org/10.1002/qj.3183

Fowler, A. M., Simonin, D., & Waller, J. A. (2020). Measuring theoretical and actual observation influence in the met office UKV: Application to doppler radial winds. *Geophysical Research Letters*, *47*(24), e2020GL091110. https://doi.org/10.1029/2020GL091110

Golub, G. H., & Van Loan, C. F. (1996). *Matrix computations*. Johns Hopkins University Press.

Greybush, S. J., Kalnay, E., Miyoshi, T., Ide, K., & Hunt, B. R. (2011). Balance and ensemble Kalman filter localization techniques. *Monthly Weather Review*, *139*(2), 511–522. https://doi.org/10.1175/2010MWR3328.1

Hamill, T. M., Whitaker, J. S., & Snyder, C. (2001). Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Monthly Weather Review*, *129*(11), 2776–2790. https://doi.org/10.1175/1520-0493(2001)129⟨2776:DDFOBE⟩2.0.CO;2

Hotta, D., & Ota, Y. (2021). Why does EnKF suffer from analysis overconfidence? An insight into exploiting the ever-increasing volume of observations. *Quarterly Journal of the Royal Meteorological Society*, *147*(735), 1258–1277. https://doi.org/10.1002/qj.3970

Houtekamer, P. L., & Mitchell, H. L. (2001). A sequential ensemble Kalman filter for atmospheric data assimilation. *Monthly Weather Review*, *129*(1), 123–137. https://doi.org/10.1175/1520-0493(2001)129⟨0123:ASEKFF⟩2.0.CO;2

Houtekamer, P. L., & Zhang, F. (2016). Review of the ensemble Kalman filter for atmospheric data assimilation. *Monthly Weather Review*, *144*(12), 4489–4532. https://doi.org/10.1175/MWR-D-15-0440.1

Hu, G. (2025). DADFS: A framework for idealised data assimilation and degrees of freedom for signal experiments. https://doi.org/10.5281/zenodo.15388900

Hu, G., & Dance, S. L. (2024). Sampling and misspecification errors in the estimation of observation-error covariance matrices using observation-minus-background and observation-minus-analysis statistics. *Quarterly Journal of the Royal Meteorological Society*, *150*, 3052–3077. https://doi.org/10.1002/qj.4750

Hu, G., Dance, S. L., Bannister, R. N., Chipilski, H. G., Guillet, O., Macpherson, B., et al. (2023). Progress, challenges, and future steps in data assimilation for convection-permitting numerical weather prediction: Report on the virtual meeting held on 10 and 12 November 2021. *Atmospheric Science Letters*, *24*(1), e1130. https://doi.org/10.1002/asl.1130

Hu, G., Dance, S. L., Fowler, A., Simonin, D., Waller, J., Auligne, T., et al. (2025). On methods for assessment of the value of observations in convection-permitting data assimilation and numerical weather forecasting. *Quarterly Journal of the Royal Meteorological Society*, *151*(768), e4933. https://doi.org/10.1002/qj.4933

Hunt, B. R., Kostelich, E. J., & Szunyogh, I. (2007). Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Physica D: Nonlinear Phenomena*, *230*(1), 112–126. https://doi.org/10.1016/j.physd.2006.11.008

Janjić, T., Nerger, L., Albertella, A., Schröter, J., & Skachko, S. (2011). On Domain localization in ensemble-based Kalman filter algorithms. *Monthly Weather Review*, *139*(7), 2046–2060. https://doi.org/10.1175/2011MWR3552.1

Kalnay, E., Ota, Y., Miyoshi, T., & Liu, J. (2012). A simpler formulation of forecast sensitivity to observations: Application to ensemble Kalman filters. *Tellus A: Dynamic Meteorology and Oceanography*, *64*(1), 18462. https://doi.org/10.3402/tellusa.v64i0.18462

Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, *88*(2), 365–411. https://doi.org/10.1016/S0047-259X(03)00096-4

Liu, J., Kalnay, E., Miyoshi, T., & Cardinali, C. (2009). Analysis sensitivity calculation in an ensemble Kalman filter. *Quarterly Journal of the Royal Meteorological Society*, *135*(644), 1842–1851. https://doi.org/10.1002/qj.511

Livings, D. M., Dance, S. L., & Nichols, N. K. (2008). Unbiased ensemble square root filters. *Physica D: Nonlinear Phenomena*, *237*(8), 1021–1028. https://doi.org/10.1016/j.physd.2008.01.005

Lorenc, A. C., Ballard, S. P., Bell, R. S., Ingleby, N. B., Andrews, P. L. F., Barker, D. M., et al. (2000). The Met. Office global three-dimensional variational data assimilation scheme. *Quarterly Journal of the Royal Meteorological Society*, *126*(570), 2991–3012. https://doi.org/10.1002/qj.49712657002

Lupu, C., Gauthier, P., & Laroche, S. (2011). Evaluation of the impact of observations on analyses in 3D- and 4D-Var based on information content. *Monthly Weather Review*, *139*(3), 726–737. https://doi.org/10.1175/2010MWR3404.1

Mitchell, H. L., & Houtekamer, P. L. (2000). An adaptive ensemble Kalman filter. *Monthly Weather Review*, *128*(2), 416–433. https://doi.org/10.1175/1520-0493(2000)128⟨0416:AAEKF⟩2.0.CO;2

Nichols, N. K. (2010). Mathematical concepts of data assimilation. In W. Lahoz, B. Khattatov, & R. Menard (Eds.), *Data assimilation: Making sense of observations* (pp. 13–39). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-74703-1_2

Ota, Y., Derber, J. C., Kalnay, E., & Miyoshi, T. (2013). Ensemble-based observation impact estimates using the NCEP GFS. *Tellus A: Dynamic Meteorology and Oceanography*, *65*(1), 20038. https://doi.org/10.3402/tellusa.v65i0.20038

Rabier, F., Järvinen, H., Klinker, E., Mahfouf, J.-F., & Simmons, A. (2000). The ECMWF operational implementation of four-dimensional variational assimilation. I: Experimental results with simplified physics. *Quarterly Journal of the Royal Meteorological Society*, *126*(564), 1143–1170. https://doi.org/10.1002/qj.49712656415

Rawlins, F., Ballard, S. P., Bovis, K. J., Clayton, A. M., Li, D., Inverarity, G. W., et al. (2007). The Met Office global four-dimensional variational data assimilation scheme. *Quarterly Journal of the Royal Meteorological Society*, *133*(623), 347–362. https://doi.org/10.1002/qj.32

Rodgers, C. D. (1998). Information content and optimisation of high spectral resolution remote measurements. *Advances in Space Research*, *21*(3), 361–367. https://doi.org/10.1016/S0273-1177(97)00915-0

Rodgers, C. D. (2000). Inverse methods for atmospheric sounding. *World Scientific*. https://doi.org/10.1142/3171

Schraff, C., Reich, H., Rhodin, A., Schomburg, A., Stephan, K., Periáñez, A., & Potthast, R. (2016). Kilometre-scale ensemble data assimilation for the COSMO model (KENDA). *Quarterly Journal of the Royal Meteorological Society*, *142*(696), 1453–1472. https://doi.org/10.1002/qj.2748

Stiller, O. (2022). New impact diagnostics for cross-validation of different observation types. *Quarterly Journal of the Royal Meteorological Society*, *148*(747), 2853–2876. https://doi.org/10.1002/qj.4339

Tabeart, J. M., Dance, S. L., Haben, S. A., Lawless, A. S., Nichols, N. K., & Waller, J. A. (2018). The conditioning of least-squares problems in variational data assimilation. *Numerical Linear Algebra with Applications*, *25*(5), e2165. https://doi.org/10.1002/nla.2165

Todling, R., Cohn, S. E., & Sivakumaran, N. S. (1998). Suboptimal schemes for retrospective data assimilation based on the fixed-lag Kalman Smoother. *Monthly Weather Review*, *126*(8), 2274–2286. https://doi.org/10.1175/1520-0493(1998)126⟨2274:SSFRDA⟩2.0.CO;2

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Vural, J., Merker, C., Löffler, M., Leuenberger, D., Schraff, C., Stiller, O., et al. (2024). Improving the representation of the atmospheric boundary layer by direct assimilation of ground-based microwave radiometer observations. *Quarterly Journal of the Royal Meteorological Society*, *150*(759), 1012–1028. https://doi.org/10.1002/qj.4634

Wahba, G., Johnson, D. R., Gao, F., & Gong, J. (1995). Adaptive tuning of numerical weather prediction models: Randomized GCV in Three- and four-dimensional data assimilation. *Monthly Weather Review*, *123*(11), 3358–3370. https://doi.org/10.1175/1520-0493(1995)123⟨3358:ATONWP⟩2.0.CO;2

Whitaker, J. S., & Hamill, T. M. (2002). Ensemble data assimilation without perturbed observations. *Monthly Weather Review*, *130*(7), 1913–1924. https://doi.org/10.1175/1520-0493(2002)130⟨1913:EDAWPO⟩2.0.CO;2

Whitaker, J. S., & Hamill, T. M. (2012). Evaluating methods to account for System errors in ensemble data assimilation. *Monthly Weather Review*, *140*(9), 3078–3089. https://doi.org/10.1175/MWR-D-11-00276.1

Zhang, F., Snyder, C., & Sun, J. (2004). Impacts of initial estimate and observation availability on convective-scale data assimilation with an ensemble Kalman filter. *Monthly Weather Review*, *132*(5), 1238–1253. https://doi.org/10.1175/1520-0493(2004)132⟨1238:IOIEAO⟩2.0.CO;2