https://doi.org/10.5194/egusphere-2022-996
Preprint. Discussion started: 10 October 2022

How far can the error estimation problem in data assimilation be closed by collocated data?

Annika Vogel[1,2] and Richard Ménard[1]

[1]Air Quality Research Division, Environment and Climate Change Canada (ECCC), Dorval - QC, Canada
[2]Rhenish Institute for Environmental Research (RIU) at the University of Cologne, Cologne, Germany

Correspondence: Annika Vogel (annika.vogel@ec.gc.ca)

Abstract. Accurate specification of error statistics required for data assimilation remains an ongoing challenge, partly because their estimation is an ill-posed problem that requires statistical assumptions. Even with the common assumption that background and observation errors are uncorrelated, the problem remains underdetermined. One natural question that could arise is: Can the increasing amount of overlapping observations or other datasets help to reduce the total number of statistical assumptions, or do they introduce more statistical unknowns? In order to answer this question, this paper provides a conceptual view on the statistical error estimation problem for multiple collocated datasets, including a generalized mathematical formulation, an exemplary demonstration with synthetic data as well as a formulation of the minimal and optimal conditions to solve the problem. It is demonstrated that the required number of statistical assumptions increases linearly with the number of datasets. However the number of error statistics that can be estimated increases quadratically, allowing for an estimation of an increasing number of error cross-statistics between datasets for more than three datasets. The presented optimal estimation of full error covariance and cross-covariance matrices between dataset does not accumulate uncertainties of assumptions among estimations of subsequent error statistics.

1 Introduction

Accurate specification of error statistics used for data assimilation has been an ongoing challenge. It is known that the accuracies of both, background and observation error covariances have a strong impact on the performance of atmospheric data assimilation (e.g., Daley, 1992a, b; Mitchell and Houtekamer, 2000; Desroziers et al., 2005; Li et al., 2009). A number of approaches to estimate optimal error statistics make use of innovations, i.e. the differences between observation and background states in observation space (Tandeo et al., 2020), but the error estimation problem remains underdetermined. Different approaches exist which aim at closing the error estimation problem, all of which rely on various assumptions. For example, error variances and correlations were estimated a-posteriori by Tangborn et al. (2002); Ménard and Deshaies-Jacques (2018); Voshtani et al. (2022) based on cross-validation of the analysis with independent observations withheld from the assimilation. However, these a-posteriori methods require an iterative calculation of the analysis and the global minimization criterion provides only spatial-mean estimates of optimal error statistics. In recent years, the amount of available datasets has increased

https://doi.org/10.5194/egusphere-2022-996
Preprint. Discussion started: 10 October 2022

rapidly, including overlapping or collocated observations from several measurements systems. This arises the question if multiple overlapping datasets can be used to estimate full spatial fields of optimal error statistics a-priori.

Outside of the field of data assimilation, two different methods were developed that allow for a statistically optimal estimation of scalar error variances for fully collocated datasets. Although being similar, these two methods were developed independently from each other in different scientific fields. One method, called the three cornered hat (3CH) method, is based on Grubbs (1948) and Gray and Allan (1974) who developed an estimation method for error variances of three datasets based on their innovations. This method is widely used in physics since decades, but and has only recently be exploited in meteorology (e.g., Anthes and Rieckh, 2018; Rieckh et al., 2021; Kren and Anthes, 2021; Xu and Zou, 2021). Nielsen et al. (2022) were the first using the generalized 3CH (G3CH) method to estimate full error covariances matrices. Todling et al. (2022) used a modification of the G3CH method to estimate the observation error covariance matrix in a data assimilation framework. However, because they use the generated analysis as third dataset, this modification does not provide a closed problem which is required to obtain optimal error statistics.

Independent from these developments, Stoffelen (1998) used three collocated datasets for multiplicative calibration w.r.t. each other. Following this idea, the triple collocation (TC) method became a well-known tool to estimate scalar error variances from innovation statistics in the fields of hydrology and oceanography (e.g., Scipal et al., 2008; McColl et al., 2014; Sjoberg et al., 2021). Up to now, only scalar error variances estimated with the TC method are rarely used in data assimilation frameworks (e.g., Crow and van den Berg, 2010; Crow and Yilmaz, 2014). The 3CH and TC methods use different error models leading to slightly different assumptions and formulations of error statistics. A detailed description, comparison and evaluation of the two methods is given in Sjoberg et al. (2021). Both methods have in common that they require fully spatio-temporally collocated datasets with random errors. These errors are assumed to be independent among the realizations of each dataset with common error statistics across all realizations (e.g., Zwieback et al., 2012; Su et al., 2014). In addition, error statistics of the three datasets are assumed to be pairwise independent, which is the most critical assumption of these methods (Pan et al., 2015; Sjoberg et al., 2021).

While the estimation of three error variances is well-established since decades, recent developments propose different approaches to extend the method to a larger number of datasets. As observed e.g. by Su et al. (2014); Pan et al. (2015); Vogelzang and Stoffelen (2021), the problem of error variance estimation from pairwise innovations becomes overdetermined for more than three datasets. Su et al. (2014); Anthes and Rieckh (2018); Rieckh et al. (2021) averaged all possible solutions of each error variance which reduces the sensitivity of the error estimates to inaccurate assumptions. Pan et al. (2015) clustered their datasets into structual groups and performed a two-step estimation of the in-group errors and the mean errors of each group, which were assumed to be independent. Zwieback et al. (2012) were the first proposing the additionally estimation of the error cross-variances between two selected datasets instead of solving an overdetermined system. This extended collocation (EC) method was applied to scalar soil moisture datasets by Gruber et al. (2016) who estimated one cross-variance in addition to the error variances of four datasets. Also for four datasets, Vogelzang and Stoffelen (2021) demonstrated the ability to estimate two cross-variances in addition to the error variances. They observed that the problem can not be solved for all possible

https://doi.org/10.5194/egusphere-2022-996
Preprint. Discussion started: 10 October 2022

combinations of cross-variances to be estimated. However, their approach failed for five dataset due to a missing generalized condition which is required to solve the problem.

This demonstrates that the different approaches available for more than three datasets provide only an incomplete picture of the problem, were each approach is tailored to the specific conditions of the respective application. Aiming for a more general analysis, this paper approaches the problem from a conceptual point-of-view. The main questions to be answered are: How many error statistics can be extracted from innovation statistics between multiple collocated datasets? How many statistics remain to be assumed? How do inaccuracies in assumed error statistics affect different formulations of the estimated error statistics? And what are the minimal and optimal conditions to solve the problem?

In order to answer these questions, the general framework of the estimation problem which builds the basis for the remaining sections is introduced in Sect. 2. It provides a conceptual analysis of the general problem w.r.t. the number of knowns and unknonws and the minimum number of assumptions required. Based on this, the mathematical formulation for non-scalar error matrices is formulated in Sect. 3 and Sect. 4, respectively. While the exact formulations in Sect. 3 remain underdetermined in real applications, approximative formulations which provide a closed system of equations are derived in Sect. 4. Some relations presented in these two sections were already formulated previously for scalar problems dealing with error variances only. However, we present formulations for full covariance matrices including off-diagonal covariances between single elements of the statevector of the respective dataset, as well as cross-covariance matrices between different datasets. Overlap to previous studies is mainly restricted to the formulation for three datasets in Sect. 4.1 and noted accordingly. Based on this, Sect. 4.2 provides a new approach for the estimation of error statistics of all additional datasets which uses a minimal number of assumptions. The theoretical formulations are exemplary applied to four synthetic datasets in Sect. 5. It demonstrates the general ability to estimate full error covariances and cross-statistics as well as effects of inaccurate assumptions w.r.t different setups. The theoretical concept proposed in this study is summarized in Sect. 6. This summary aims at providing the most important results in a general context; answering the main research questions of this study without requiring the knowledge of the full mathematical theory. It includes the formulation and illustration of minimal requirements to solve the problem for an arbitrary number of datasets and provides criteria for an optimal setup of those. Finally, Sect. 7 concludes the findings and discusses consequences of using the proposed method in the context of high-dimensional data assimilation.

2 General framework

Suppose a system of I spatio-temporally collocated datasets, which may include various model forecasts, observations, analyses or any other datasets available in the same state space. The 2nd moment statistics of the random errors of this system (with respect to the truth) can be described by I error covariances w.r.t. each dataset and N_I error cross-covariances w.r.t. each pair of different datasets. In a discrete state space, (cross-)covariances are matrices and the cross-covariance of dataset A and B is the transposed of the cross-covariance of B and A (see Sect. 3.1 for an explicit definition). Considering this equivalence, the

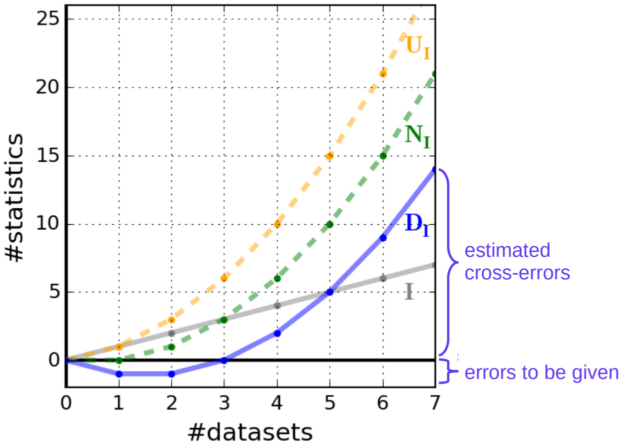

Figure 1. Relation between different numbers of statistics (covariances ans cross-covariances) as function of the number of datasets. Shown are I in solid gray (#datasets, #error covariances, #required assumptions), U_I in dashed orange (#error statistics), N_I in dashed green (#innovation covariances, #error dependencies, #estimated error statistics), and D_I in solid blue (#estimated error dependencies).

number N_I of error cross-covariances between all different pairs of datasets is:

90 $$N_I = \sum_{i=1}^{I-1} i = \frac{1}{2} \cdot I \cdot (I-1) \tag{1}$$

Thus, the total number U_I of error statistics (error covariances and cross-covariances) is:

$$U_I = N_I + I \overset{(1)}{=} \frac{1}{2} \cdot I \cdot (I+1) \tag{2}$$

While error statistics w.r.t. the truth are usually unknown in real applications, innovation covariances can be calculated from the innovations between each pair of different datasets. Because innovation statistics do not change with the order of datasets 95 in the innovation (see Sect. 3.1), the number of known statistics of the system is also given by N_I as defined in Eq. (1).

Because N_I innovation statistics are known, N_I of the U_I error statistics can be estimated and the remaining I have to be assumed in order to close the problem. The set of error statistics to be estimated can be chosen according to the specific application under some restrictions as discussed in Sect. 6.1.

In most applications of geophysical datasets like in data assimilation, the estimation of error covariances is highly crucial 100 while their error cross-covariances are usually assumed to be neglectable. Given the greater need to estimate the I error covariances, the remaining number of error cross-covariances which can be additionally estimated D_I is:

$$D_I = N_I - I \overset{(1)}{=} \frac{1}{2} \cdot I \cdot (I-3) \tag{3}$$

The relation between the number of datasets, innovation covariances, assumed- and estimated error statistics is visualized in Fig. 1. $I = 0$ represents the mathematical extension of the problem, were no error and innovation statistics are required when no dataset is considered. For less than three datasets ($0 < I < 3$), D_I is negative because the number of (known) innovation covariances is smaller than the number of (unknown) error covariances ($N_I < I$) and thus the problem is underdetermined even when all datasets are assumed to be independent (=zero error cross-covariances). As it is the case in data assimilation of two datasets ($I = 2$), additional assumptions on error statistics are required. The same holds when only one dataset is available ($I = 1$), were the error covariance of this dataset remains unknown because no innovation covariance can be formed. For three datasets ($I = 3$), D_I is zero meaning that the problem is fully determined under the assumption of independent errors ($N_I = I$, formulated in Sect. 4.1).

For more than three datasets ($I > 3$), the number of (known) innovation covariances exceeds the number of error covariances which would lead to an overdetermined problem assuming independence between all datasets. Instead of solving an overdetermined problem, the additional information can be used to calculate some error cross-covariances (formulated in Sect. 4.2). In other words, for $I > 3$ not all datasets need to be assumed to be independent; were D_I gives the number of error cross-covariances which can be estimated in addition to the error covariances from all datasets. For example, half of the error cross-covariances can be estimated for $I = 5$ $\left(\frac{D_5}{N_5} = \frac{5}{10}\right)$, while two-thirds of them can be estimated for $I = 7$ $\left(\frac{D_7}{N_7} = \frac{14}{21}\right)$. Although the relative amount of error cross-covariances which can be estimated increases with the number of datasets, an increasing number of $U_I - N_I = I$ assumptions – equal to the number of datasets – is required in order to close the problem because of $U_I > N_I \ ,\forall\, I > 0$.

3 Mathematical theory: Exact formulation

This section gives the theoretical formulation for exact statistical formulations of complete error covariance and cross-covariance matrices from fully spatio-temporally collocated datasets. Similar to the 3CH method, the errors are assumed to be random, independent among different realizations, but with common error statistics for each dataset. The notation is introduced in Sect. 3.1. While the true state and thus error statistics w.r.t. the truth are usually unknown, innovation statistics can be calculated from innovations between each pair of datasets. At the same time, innovation statistics contain information about error statistics of the datasets involved. The general formulation of this forward relation between error- and innovation statistics is given in Sect. 3.2. Based on these forward relations, inverse relations describe error statistics as function of innovation statistics. The general equations of inverse relations are given in Sect. 3.3 which result in a highly underdetermined system of equations. Closed formulations of error statistics for three and more datasets under certain assumptions will be formulated in the subsequent Sect. 4.

This first part of the mathematical theory includes the following new elements: (i) the separation of cross-statistics into a symmetric error dependency and an error asymmetry, (ii) the general formulation of exact relations between innovation- and error statistics.

3.1 Notation

Suppose I datasets, each containing R realizations of spatio-temporally collocated state vectors $\boldsymbol{x}'_{i|r}\ \forall\, i \in [1,I], \forall\, r \in [1,R]$. Without loss of generality, the following formulation uses unbiased state vectors $\boldsymbol{x}_{i|r} := \boldsymbol{x}'_{i|r} - \overline{\boldsymbol{x}_i}$ with zero mean, were the overbar indicates its expectation w.r.t. realizations.

Let $\boldsymbol{\Gamma}_{i\text{-}j;k\text{-}l}$ be the *innovation cross-covariance matrix* between dataset innovations $i-j$ and $k-l$, were each element (p,q) is given by:

$$\boldsymbol{\Gamma}_{i\text{-}j;k\text{-}l}(p,q) := \overline{\left[\boldsymbol{x}_i(p) - \boldsymbol{x}_j(p)\right] \cdot \left[\boldsymbol{x}_k(q) - \boldsymbol{x}_l(q)\right]} \tag{4}$$

and the *error cross-covariance matrix* $\mathbf{X}_{\tilde{i};\tilde{j}}$ between the errors of two datasets i and j w.r.t. the true state $\boldsymbol{x}_T$:

$$\mathbf{X}_{\tilde{i};\tilde{j}}(p,q) := \overline{\left[\boldsymbol{x}_i(p) - \boldsymbol{x}_T(p)\right] \cdot \left[\boldsymbol{x}_j(q) - \boldsymbol{x}_T(q)\right]} \tag{5}$$

were the tilde above an dataset index indicates its deviation from the truth.

In the symmetric case, each element (p,q) of the *innovation covariance matrix* of $i-j$, is given by:

$$\boldsymbol{\Gamma}_{i\text{-}j}(p,q) := \boldsymbol{\Gamma}_{i\text{-}j;i\text{-}j}(p,q) \overset{(4)}{=} \overline{\left[\boldsymbol{x}_i(p) - \boldsymbol{x}_j(p)\right] \cdot \left[\boldsymbol{x}_i(q) - \boldsymbol{x}_j(q)\right]} \tag{6}$$

and the *error covariance matrix* $\mathbf{C}_{\tilde{i}}$ of a dataset i w.r.t. the true state $\boldsymbol{x}_T$:

$$\mathbf{C}_{\tilde{i}}(p,q) := \mathbf{X}_{\tilde{i};\tilde{i}}(p,q) \overset{(5)}{=} \overline{\left[\boldsymbol{x}_i(p) - \boldsymbol{x}_T(p)\right] \cdot \left[\boldsymbol{x}_i(q) - \boldsymbol{x}_T(q)\right]} \tag{7}$$

Note that innovation- and error cross-covariance matrices are generally asymmetric in the non-scalar formulation presented here, but the following relations hold for innovation- as well as similarly for error cross-covariance matrices:

$$\boldsymbol{\Gamma}_{i\text{-}j;k\text{-}l} \overset{(4)}{=} -\boldsymbol{\Gamma}_{j\text{-}i;k\text{-}l} \overset{(4)}{=} -\boldsymbol{\Gamma}_{i\text{-}j;l\text{-}k} \overset{(4)}{=} \boldsymbol{\Gamma}_{j\text{-}i;l\text{-}k} \tag{8}$$

$$\boldsymbol{\Gamma}_{i\text{-}j;k\text{-}l} \overset{(4)}{=} \left[\boldsymbol{\Gamma}_{k\text{-}l;i\text{-}j}\right]^T \tag{9}$$

$$\mathbf{X}_{\tilde{i};\tilde{j}} \overset{(5)}{=} \left[\mathbf{X}_{\tilde{j};\tilde{i}}\right]^T \tag{10}$$

The symmetric properties of innovation- and error covariances follow directly from their definition:

$$\boldsymbol{\Gamma}_{i\text{-}j} \overset{(6)}{=} \boldsymbol{\Gamma}_{j\text{-}i} \tag{11}$$

$$\left[\boldsymbol{\Gamma}_{i\text{-}j}\right]^T \overset{(6)}{=} \boldsymbol{\Gamma}_{i\text{-}j} \tag{12}$$

The sum of an (asymmetric) cross-covariance matrix and its transposed is denoted as *dependency*. For example, sum of the error cross-covariance matrices between i and j is denoted as *error dependency matrix* $\mathbf{D}_{\tilde{i};\tilde{j}}$:

$$\mathbf{D}_{\tilde{i};\tilde{j}} := \mathbf{X}_{\tilde{i};\tilde{j}} + \mathbf{X}_{\tilde{j};\tilde{i}} \tag{13}$$

https://doi.org/10.5194/egusphere-2022-996
Preprint. Discussion started: 10 October 2022

Although error cross-covariances may be asymmetric, the error dependency matrix is symmetric by definition:

$$\mathbf{D}_{\tilde{i};\tilde{j}} \overset{(13)}{=} \mathbf{X}_{\tilde{i};\tilde{j}} + \mathbf{X}_{\tilde{j};\tilde{i}} \overset{(13)}{=} \mathbf{D}_{\tilde{j};\tilde{i}} \tag{14}$$

$$\mathbf{D}_{\tilde{i};\tilde{j}} \overset{(13)}{=} \mathbf{X}_{\tilde{i};\tilde{j}} + \mathbf{X}_{\tilde{j};\tilde{i}} \overset{(10)}{=} \left[\mathbf{X}_{\tilde{j};\tilde{i}}\right]^T + \left[\mathbf{X}_{\tilde{i};\tilde{j}}\right]^T \overset{(13)}{=} \left[\mathbf{D}_{\tilde{i};\tilde{j}}\right]^T \tag{15}$$

Likewise, the sum of the innovation cross-covariance matrices between $i-j$ and $k-l$ is denoted as *innovation dependency matrix* $\mathbf{D}_{\tilde{i};\tilde{j}}$:

$$\mathbf{D}_{i\text{-}j;k\text{-}l} := \mathbf{\Gamma}_{i\text{-}j;k\text{-}l} + \mathbf{\Gamma}_{k\text{-}l;i\text{-}j} \tag{16}$$

The difference between a cross-covariance matrix and its transposed is a measure of asymmetry in the cross-covariances and is therefore denoted as *asymmetry*. For example, difference between the error cross-covariance matrices between i and j is denoted as *error asymmetry matrix* $\mathbf{Y}_{\tilde{i};\tilde{j}}$:

$$\mathbf{Y}_{\tilde{i};\tilde{j}} := \mathbf{X}_{\tilde{i};\tilde{j}} - \mathbf{X}_{\tilde{j};\tilde{i}} \tag{17}$$

Likewise, the difference between the innovation cross-covariance matrices between $i-j$ and $k-l$ is denoted as *innovation asymmetry matrix* $\mathbf{D}_{\tilde{i};\tilde{j}}$:

$$\mathbf{Y}_{i\text{-}j;k\text{-}l} := \mathbf{\Gamma}_{i\text{-}j;k\text{-}l} - \mathbf{\Gamma}_{k\text{-}l;i\text{-}j} \tag{18}$$

In practice, each index i,j,k,l may represent any geophysical dataset like model forecasts, climatologies, in-situ- or remote sensing observations, or other datasets.

3.2 Innovation statistics

For real geophysical problems, the available statistical information are (i) innovation covariance matrices of each pair of datsets and (ii) innovation cross-covariance matrices between different innovations of datsets. The forward relations of innovation covariances and innovation cross-covariances as function of error statistics are formulated in the following. For the estimation of error statistics, it is important to quantify the number of independent input statistics which determines the number of possible error estimations. Therefore, this section also includes an evaluation of the relation between innovation cross-covariances and innovation covariances in order to specify the additional information content of innovation cross-covariances.

3.2.1 Innovation covariances

Each element (p,q) of the innovation covariance matrix between two input datasets i and j can be written as function of their error statistics as follows:

$$\begin{aligned}\mathbf{\Gamma}_{i\text{-}j}(p,q) &\overset{(6)}{=} \overline{\Big\{\big[\boldsymbol{x}_i(p) - \boldsymbol{x}_T(p)\big] - \big[\boldsymbol{x}_j(p) - \boldsymbol{x}_T(p)\big]\Big\} \cdot \Big\{\big[\boldsymbol{x}_i(q) - \boldsymbol{x}_T(q)\big] - \big[\boldsymbol{x}_j(q) - \boldsymbol{x}_T(q)\big]\Big\}} \\ &\overset{(5),(7)}{=} \mathbf{C}_{\tilde{i}}(p,q) - \mathbf{X}_{\tilde{i};\tilde{j}}(p,q) - \mathbf{X}_{\tilde{j};\tilde{i}}(p,q) + \mathbf{C}_{\tilde{j}}(p,q)\end{aligned} \tag{19}$$

https://doi.org/10.5194/egusphere-2022-996
Preprint. Discussion started: 10 October 2022

Thus the complete innovation covariance matrix of $i-j$ is expressed as:

$$\mathbf{\Gamma}_{i\text{-}j} \overset{(19)}{=} \underbrace{\mathbf{C}_{\widetilde{i}} + \mathbf{C}_{\widetilde{j}}}_{\text{"independent innovation"}} \quad - \quad \underbrace{\left[\mathbf{X}_{\widetilde{i};\widetilde{j}} + \mathbf{X}_{\widetilde{j};\widetilde{i}}\right]}_{\text{"error dependency"} =: \mathbf{D}_{\widetilde{i};\widetilde{j}}} \tag{20}$$

Equation (20) is an exact formulation of the complete innovation covariance matrix of any pair of datasets $i-j$. It holds for all combinations of datasets without any further assumption like independent- or unbiased error statistics. Thus, for every pair of datasets, their innovation covariance consists of an independent innovation being the sum of their error covariances subtracted by their innovation dependency being the sum of their error cross-covariances.

Note that although the error dependency matrix is symmetric by definition, it is the sum of two error cross-covariances which are generally asymmetric and thus differ in the non-scalar formulation. In the scalar case, the two error cross-covariances reduce to their common error cross-variance and the innovation covariance reduces to the scalar formulation of the variance as e.g. in Anthes and Rieckh (2018); Sjoberg et al. (2021).

3.2.2 Innovation cross-covariances

Each element (p,q) of the innovation cross-covariance matrix between two input datasets $i-j$ and $k-l$ can be written as function of their error cross-covariances:

$$\begin{aligned}\mathbf{\Gamma}_{i\text{-}j;k\text{-}l}(p,q) &\overset{(4)}{=} \overline{\left\{\left[\boldsymbol{x}_i(p) - \boldsymbol{x}_T(p)\right] - \left[\boldsymbol{x}_j(p) - \boldsymbol{x}_T(p)\right]\right\} \cdot \left\{\left[\boldsymbol{x}_k(q) - \boldsymbol{x}_T(q)\right] - \left[\boldsymbol{x}_l(q) - \boldsymbol{x}_T(q)\right]\right\}} \\ &\overset{(5)}{=} \mathbf{X}_{\widetilde{i};\widetilde{k}}(p,q) - \mathbf{X}_{\widetilde{i};\widetilde{l}}(p,q) - \mathbf{X}_{\widetilde{j};\widetilde{k}}(p,q) + \mathbf{X}_{\widetilde{j};\widetilde{l}}(p,q)\end{aligned} \tag{21}$$

And thus the complete innovation cross-covariance matrix between $i-j$ and $k-l$:

$$\mathbf{\Gamma}_{i\text{-}j;k\text{-}l} \overset{(21)}{=} \mathbf{X}_{\widetilde{i};\widetilde{k}} - \mathbf{X}_{\widetilde{i};\widetilde{l}} - \mathbf{X}_{\widetilde{j};\widetilde{k}} + \mathbf{X}_{\widetilde{j};\widetilde{l}} \tag{22}$$

Equation (22) is a generalized form of Eq. (20) with innovations between different datasets ($i-j$; $k-l$). It consists of four error cross-covariances of the datasets involved. In contrast to the symmetric innovation covariance matrix, the innovation cross-covariance matrix may be asymmetric for asymmetric error cross-covariances.

3.2.3 Relation of innovation statistics

In the following, it is demonstrated that combinations of innovation cross-covariances contain the same statistical information as innovation covariance matrices.

For $k=i$, the innovation dependency between $i-j$ and $i-l$ becomes:

$$\begin{aligned}\mathbf{D}_{i\text{-}j;i\text{-}l} := \mathbf{\Gamma}_{i\text{-}j;i\text{-}l} + \mathbf{\Gamma}_{i\text{-}l;i\text{-}j} &\overset{(21)}{=} \mathbf{C}_{\widetilde{i}} - \mathbf{X}_{\widetilde{i};\widetilde{l}} - \mathbf{X}_{\widetilde{j};\widetilde{i}} + \mathbf{X}_{\widetilde{j};\widetilde{l}} + \mathbf{C}_{\widetilde{i}} - \mathbf{X}_{\widetilde{i};\widetilde{j}} - \mathbf{X}_{\widetilde{l};\widetilde{i}} + \mathbf{X}_{\widetilde{l};\widetilde{j}} \\ &\overset{(13)}{=} 2\,\mathbf{C}_{\widetilde{i}} - \mathbf{D}_{\widetilde{i};\widetilde{l}} - \mathbf{D}_{\widetilde{j};\widetilde{i}} + \mathbf{D}_{\widetilde{j};\widetilde{l}} \quad + \left[\mathbf{C}_{\widetilde{j}} + \mathbf{C}_{\widetilde{j}}\right] - \left[\mathbf{C}_{\widetilde{j}} + \mathbf{C}_{\widetilde{j}}\right] = \mathbf{C}_{\widetilde{i}} + \mathbf{C}_{\widetilde{l}} - \mathbf{D}_{\widetilde{i};\widetilde{l}} + \mathbf{C}_{\widetilde{j}} + \mathbf{C}_{\widetilde{i}} - \mathbf{D}_{\widetilde{j};\widetilde{i}} - \mathbf{C}_{\widetilde{j}} - \mathbf{C}_{\widetilde{l}} + \mathbf{D}_{\widetilde{j};\widetilde{l}} \\ \Longrightarrow \quad \mathbf{\Gamma}_{i\text{-}j;i\text{-}l} + \mathbf{\Gamma}_{i\text{-}l;i\text{-}j} &\overset{(20)}{=} \mathbf{\Gamma}_{i\text{-}l} + \mathbf{\Gamma}_{j\text{-}i} - \mathbf{\Gamma}_{j\text{-}l}\end{aligned} \tag{23}$$

https://doi.org/10.5194/egusphere-2022-996
Preprint. Discussion started: 10 October 2022

The relation between innovation covariances and innovation cross-covariances in Eq. (23) is exact and holds for all datasets without any further assumption. In case of symmetric innovation cross-covariances $\left(\mathbf{\Gamma}_{i\text{-}j;i\text{-}l} = \mathbf{\Gamma}_{i\text{-}l;i\text{-}j} \overset{(23)}{=} \frac{1}{2}\left[\mathbf{\Gamma}_{i\text{-}l} + \mathbf{\Gamma}_{j\text{-}i} - \mathbf{\Gamma}_{j\text{-}l}\right]\right)$, the innovation cross-covariance matrices are fully determined by the symmetric innovation covariances.

In the general asymmetric case, Eq. (23) can be rewritten as:

$$\begin{aligned}
&\mathbf{\Gamma}_{i\text{-}l} + \mathbf{\Gamma}_{j\text{-}i} - \mathbf{\Gamma}_{j\text{-}l} \overset{(23)}{=} \mathbf{\Gamma}_{i\text{-}j;i\text{-}l} + \mathbf{\Gamma}_{i\text{-}l;i\text{-}j} \overset{(18)}{=} \mathbf{\Gamma}_{i\text{-}j;i\text{-}l} + \left[\mathbf{\Gamma}_{i\text{-}j;i\text{-}l} - \mathbf{Y}_{i\text{-}j;i\text{-}l}\right] \\
&\iff \quad \mathbf{\Gamma}_{i\text{-}j;i\text{-}l} = \frac{1}{2}\left[\mathbf{\Gamma}_{i\text{-}l} + \mathbf{\Gamma}_{j\text{-}i} - \mathbf{\Gamma}_{j\text{-}l}\right] + \frac{1}{2}\mathbf{Y}_{i\text{-}j;i\text{-}l} \qquad (24)
\end{aligned}$$

Equation (24) shows that each individual innovation cross-covariance consists of a symmetric contribution including innovation covariances between the datasets and an asymmetric contribution being half of the related innovation asymmetry matrix. Thus, innovation cross-covariances may only provide additional information on asymmetries of error statistics, but not on symmetric statistics (like error covariances).

3.3 Exact error statistics

As an extension to previous work, this section provides generalized formulations of error covariances, -cross-covariances, and -dependencies in matrix form. Note that the general formulations presented here do not provide a closed system of equations which can be solved in real applications. They serve as basis for the approximative solutions which are formulated in the subsequent section.

3.3.1 Error statistics from innovation covariances

Equation (20) shows that each innovation covariance matrix can be expressed by the error covariances of the two datasets involved and their error dependency. The goal is to find an inverse formulation of an error covariance matrix as function of innovation covariances which does not include other (unknown) error covariances matrices. This is achieved by combining the formulations of three innovations $\mathbf{\Gamma}_{i\text{-}j}$, $\mathbf{\Gamma}_{j\text{-}k}$, and $\mathbf{\Gamma}_{k\text{-}i}$ between the same three datasets i, j, and k which eliminates a single error covariance:

$$\begin{aligned}
&\mathbf{C}_{\widetilde{i}} \overset{(20)_{ij}}{=} \mathbf{\Gamma}_{i\text{-}j} + \mathbf{D}_{\widetilde{i};\widetilde{j}} - \mathbf{C}_{\widetilde{j}} \qquad (25) \\
&\overset{(20)_{jk}}{=} \mathbf{\Gamma}_{i\text{-}j} + \mathbf{D}_{\widetilde{i};\widetilde{j}} - \mathbf{\Gamma}_{j\text{-}k} - \mathbf{D}_{\widetilde{j};\widetilde{k}} + \mathbf{C}_{\widetilde{k}} \overset{(20)_{ki}}{=} \mathbf{\Gamma}_{i\text{-}j} + \mathbf{D}_{\widetilde{i};\widetilde{j}} - \mathbf{\Gamma}_{j\text{-}k} - \mathbf{D}_{\widetilde{j};\widetilde{k}} + \mathbf{\Gamma}_{k\text{-}i} + \mathbf{D}_{\widetilde{k};\widetilde{i}} - \mathbf{C}_{\widetilde{i}} \\
&\iff \quad \mathbf{C}_{\widetilde{i}} = \frac{1}{2}\Bigg[\underbrace{\mathbf{\Gamma}_{i\text{-}j} + \mathbf{\Gamma}_{k\text{-}i} - \mathbf{\Gamma}_{j\text{-}k}}_{\text{"independent contribution"}} + \underbrace{\mathbf{D}_{\widetilde{i};\widetilde{j}} + \mathbf{D}_{\widetilde{k};\widetilde{i}} - \mathbf{D}_{\widetilde{j};\widetilde{k}}}_{\text{"dependent contribution"}}\Bigg] \qquad (26)
\end{aligned}$$

Equation (26) provides a general formulation of error covariances as function of innovation covariances and error dependencies. It holds for all combinations of datasets without any further assumption (e.g. independence). Thus, each error covariance can be formulated as sum of an independent contribution of three innovation covariances w.r.t. any pair of other datasets and an dependent contribution of the three related error dependencies. Note that while the independent contribution can be calculated from innovation statistics between input datsets, the dependent contribution is generally unknown in real applications.

Given I datasets, the total number of different formulations of each error covariance by Eq. (26) is determined by the number of different pairs of the other datasets which is $\sum_{i=1}^{I-2} i = \frac{1}{2}(I-1)(I-2)$ (see also Sjoberg et al., 2021). The scalar equivalent of Eq. (26) were the dependency matrices reduce to twice the error cross-variances has been formulated previously in the 3CH method e.g. in Anthes and Rieckh (2018); Sjoberg et al. (2021). Very recently, the full matrix form was used by Nielsen et al. (2022); Todling et al. (2022).

A formulation of each individual error dependency matrix as function of the error covariances of the two datasets and their innovation covariance results directly from Eq. (20):

$$\mathbf{D}_{\tilde{i};\tilde{j}} \overset{(20)_{ij}}{=} \mathbf{C}_{\tilde{i}} + \mathbf{C}_{\tilde{j}} - \mathbf{\Gamma}_{i\text{-}j} \tag{27}$$

Being a symmetric matrix, innovation covariances cannot provide information on error asymmetries and thus on asymmetric components of error cross-covariances. Only the symmetric component of error cross-covariances could be estimated from half the error dependency which is equivalent to a zero error asymmetry matrix:

$$\mathbf{D}_{\tilde{i};\tilde{j}} + \mathbf{Y}_{\tilde{i};\tilde{j}} \overset{(13),(17)}{=} \left[\mathbf{X}_{\tilde{i};\tilde{j}} \cancel{+\mathbf{X}_{\tilde{j};\tilde{i}}}\right] + \left[\mathbf{X}_{\tilde{i};\tilde{j}} \cancel{-\mathbf{X}_{\tilde{j};\tilde{i}}}\right] \qquad \Longleftrightarrow \qquad \mathbf{X}_{\tilde{i};\tilde{j}} = \frac{1}{2}\left[\mathbf{D}_{\tilde{i};\tilde{j}} + \mathbf{Y}_{\tilde{i};\tilde{j}}\right] \tag{28}$$

3.3.2 Error statistics from innovation cross-covariances

The general forward formulation of innovation cross-covariances in Eq. (22) consists of error cross-covariances of the four datasets involved. Setting for example $k = i$, provides an inverse formulation of error covariances of i:

$$\mathbf{\Gamma}_{i\text{-}j;i\text{-}l} \overset{(22)_{k=i}}{=} \mathbf{C}_{\tilde{i}} - \mathbf{X}_{\tilde{i};\tilde{l}} - \mathbf{X}_{\tilde{j};\tilde{i}} + \mathbf{X}_{\tilde{j};\tilde{l}} \qquad \Longleftrightarrow \qquad \mathbf{C}_{\tilde{i}} = \mathbf{\Gamma}_{i\text{-}j;i\text{-}l} + \mathbf{X}_{\tilde{i};\tilde{l}} + \mathbf{X}_{\tilde{j};\tilde{i}} - \mathbf{X}_{\tilde{j};\tilde{l}} \tag{29}$$

The scalar formulation of Eq. (29) was previously formulated in Zwieback et al. (2012).

Similarly to the formulation from innovation covariances, the number of formulations of each error covariance from different pairs of other datasets in Eq. (29) is $\sum_{i=1}^{I-2} i = \frac{1}{2}(I-1)(I-2)$. In addition, there are four possibilities to write each error covariances from the same pairs of other datasets using the relations of innovation cross-covariances in Eq. (8). Each of the four possibilities results from setting one pair of datsets in definition of innovation cross-covariances in Eq. (22) equal.

Two of the error cross-covariances in Eq. (29) can be replaced by a formulation of another error covariance, using:

$$\mathbf{C}_{\tilde{j}} \overset{(29)_{jijl}}{=} \mathbf{\Gamma}_{j\text{-}i;j\text{-}l} + \mathbf{X}_{\tilde{j};\tilde{l}} + \mathbf{X}_{\tilde{i};\tilde{j}} - \mathbf{X}_{\tilde{i};\tilde{l}} \qquad \Longleftrightarrow \qquad \mathbf{X}_{\tilde{i};\tilde{l}} - \mathbf{X}_{\tilde{j};\tilde{l}} = \mathbf{\Gamma}_{j\text{-}i;j\text{-}l} + \mathbf{X}_{\tilde{i};\tilde{j}} - \mathbf{C}_{\tilde{j}} \tag{30}$$

With this, Eq. (29) becomes:

$$\mathbf{C}_{\tilde{i}} \overset{(30)}{=} \mathbf{\Gamma}_{i\text{-}j;i\text{-}l} + \mathbf{\Gamma}_{j\text{-}i;j\text{-}l} - \mathbf{C}_{\tilde{j}} + \mathbf{X}_{\tilde{i};\tilde{j}} + \mathbf{X}_{\tilde{j};\tilde{i}} \overset{(13)}{=} \mathbf{\Gamma}_{i\text{-}j;i\text{-}l} + \mathbf{\Gamma}_{j\text{-}i;j\text{-}l} - \mathbf{C}_{\tilde{j}} + \mathbf{D}_{\tilde{i};\tilde{j}} \tag{31}$$

Because the innovation cross-covariances can be rewritten as:

$$\mathbf{\Gamma}_{i\text{-}j;i\text{-}l} + \mathbf{\Gamma}_{j\text{-}i;j\text{-}l} \overset{(22)}{=} \mathbf{C}_{\tilde{i}} \cancel{-\mathbf{X}_{\tilde{i};\tilde{l}}} - \mathbf{X}_{\tilde{j};\tilde{i}} \cancel{+\mathbf{X}_{\tilde{j};\tilde{l}}} + \mathbf{C}_{\tilde{j}} \cancel{-\mathbf{X}_{\tilde{j};\tilde{l}}} - \mathbf{X}_{\tilde{i};\tilde{j}} \cancel{+\mathbf{X}_{\tilde{i};\tilde{l}}} \overset{(13)}{=} \mathbf{C}_{\tilde{i}} + \mathbf{C}_{\tilde{j}} - \mathbf{D}_{\tilde{i};\tilde{j}} \overset{(20)}{=} \mathbf{\Gamma}_{i\text{-}j} \tag{32}$$

the formulation of error covariances based on innovation cross-covariances in Eq. (31) is symmetric and equivalent to the formulation based on innovation covariances from Eq. (25).

The forward formulation of innovation cross-covariances does not allow for an elimination of one single error cross-covariance even when multiple equations are combined. One formulation of an error cross-covariance matrix as function of innovation cross-covariances results directly from the forward relation:

$$\mathbf{X}_{\widetilde{j};\widetilde{l}} \overset{(29)}{=} \mathbf{\Gamma}_{i\text{-}j;i\text{-}l} - \mathbf{C}_{\widetilde{i}} + \mathbf{X}_{\widetilde{i};\widetilde{l}} + \mathbf{X}_{\widetilde{j};\widetilde{i}} \tag{33}$$

Note that multiple relations like Eq. (33) can be formulated analog to the formulation of error covariances from innovation cross-covariances, which are all equivalent in the exact formulation.

Any of the formulations of error cross-covariances can also be used for a formulation of the error dependency matrix $\mathbf{D}_{\widetilde{j};\widetilde{l}}\Big|_{\text{cross}}$ which is equivalent to the formulation based on innovation covariances $\mathbf{D}_{\widetilde{j};\widetilde{l}}\Big|_{\text{covar}}$:

$$\begin{aligned}
\mathbf{D}_{\widetilde{j};\widetilde{l}}\Big|_{\text{cross}} &\overset{(13)}{=} \mathbf{X}_{\widetilde{j};\widetilde{l}} + \mathbf{X}_{\widetilde{l};\widetilde{j}} \overset{(33)}{=} \mathbf{\Gamma}_{j\text{-}i;l\text{-}i} - \mathbf{C}_{\widetilde{i}} + \mathbf{X}_{\widetilde{j};\widetilde{i}} + \mathbf{X}_{\widetilde{i};\widetilde{l}} + \mathbf{\Gamma}_{l\text{-}i;j\text{-}i} - \mathbf{C}_{\widetilde{i}} + \mathbf{X}_{\widetilde{i};\widetilde{j}} + \mathbf{X}_{\widetilde{l};\widetilde{i}} \\
&\overset{(13)}{=} \mathbf{\Gamma}_{j\text{-}i;l\text{-}i} + \mathbf{\Gamma}_{l\text{-}i;j\text{-}i} - 2\,\mathbf{C}_{\widetilde{i}} + \mathbf{D}_{\widetilde{i};\widetilde{j}} + \mathbf{D}_{\widetilde{i};\widetilde{l}} \overset{(23)}{=} \mathbf{\Gamma}_{i\text{-}j} + \mathbf{D}_{\widetilde{i};\widetilde{j}} + \mathbf{\Gamma}_{i\text{-}l} + \mathbf{D}_{\widetilde{i};\widetilde{l}} - \mathbf{\Gamma}_{j\text{-}l} - 2\,\mathbf{C}_{\widetilde{i}} \\
&\overset{(20)}{=} \cancel{\mathbf{C}_{\widetilde{i}}} + \mathbf{C}_{\widetilde{j}} \cancel{+ \mathbf{C}_{\widetilde{i}}} + \mathbf{C}_{\widetilde{l}} - \mathbf{\Gamma}_{j\text{-}l} \cancel{- 2\,\mathbf{C}_{\widetilde{i}}} = \mathbf{C}_{\widetilde{j}} + \mathbf{C}_{\widetilde{l}} - \mathbf{\Gamma}_{j\text{-}l} \quad \overset{(27)}{=} \mathbf{D}_{\widetilde{j};\widetilde{l}}\Big|_{\text{covar}}
\end{aligned} \tag{34}$$

The equivalence demonstrates that the exact formulations of error statistics from innovation covariances and -cross-covariances are consistent to each other.

4 Mathematical theory: Approximative formulation

Based on the exact formulations in Sect. 3 which remain underdetermined in real applications, this section provides approximative formulations for three and more datasets which provide a closed system of equations. Section 4.1 describes the long-known closure of the system for three datasets, but for full covariance matrices. An extension for any additional dataset $I > 3$ using a minimal number of assumptions is introduced in Sect. 4.2. It includes the estimation of additional error covariances and some error cross-statistics to estimate a maximum amount of error statistics.

In addition to the optimal extension to more than three datasets, this second part of the mathematical theory includes the following new elements: (i) the determination of uncertainties caused by assumed error statistics, and (ii) the analysis of differences between error estimates from innovation covariance and cross-covariance matrices.

4.1 Approximation for three datasets

As demonstrated in Sect. 2, at least three collocated datasets are required to estimate all error covariances ($U_I \geq 0$). For three datasets ($I = 3$), three innovation covariances ($N_3 = 3$) can be calculated between each pair of datasets. At the same time, there are six unknown error statistics ($U_3 = 6$): three error covariances and three error cross-statistics (cross-covariances or dependencies). Thus, the problem is under-determined and three error statistics ($U_3 - N_3 = 3$) have to be assumed in order to close the system. The most common approach, which is also used in 3CH and TC methods, is to assume zero error cross-statistics between all pairs of datasets: $\mathbf{X}_{\widetilde{i};\widetilde{j}} = 0 \Leftrightarrow \mathbf{D}_{\widetilde{i};\widetilde{j}} = 0 \;, \forall\; i,j \in [1,3],\; j \neq i$. The approximation of the three error covariances can also be formulated in a Hilbert space which allows for an illustrative geometric interpretation as in Pan et al. (2015) (their Fig. 1). Because the assumption of zero cross-covariance equals zero error dependency, it is denoted as "assumption of independence" or "independent assumption" thereafter.

The independent assumption is consistent to the innovation covariance consistency in data assimilation, were the innovation covariance between two datasets is assumed to be equal to the sum of their error covariances in the formulation of the analysis (e.g., Daley, 1992b; Ménard, 2016):

$$\mathbf{\Gamma}_{i\text{-}j} \underset{\{in\}}{\overset{(20)}{\approx}} \mathbf{C}_{\widetilde{i}} + \mathbf{C}_{\widetilde{j}} \tag{35}$$

Here, " $\underset{\{in\}}{\approx}$ " indicates the assumption of independence between the two datasets.

Because all error cross-statistics need to be assumed in this setup, approximations of these cross-covariances and dependencies only reproduces the initially assumed statistics and do not provide any new information.

4.1.1 Error covariance estimates

Assuming independent error statistics between all three datasets, or similarly that error dependencies are neglectable compared to innovation covariances $\mathbf{D}_{\widetilde{i};\widetilde{j}} \ll \mathbf{\Gamma}_{i\text{-}j} \; \forall\, j \neq i$, gives an estimate of each error covariance matrix as function of three innovation covariances:

$$\mathbf{C}_{\widetilde{i}} \underset{\{in3\}}{\overset{(26)}{\approx}} \frac{1}{2}\left[\mathbf{\Gamma}_{i\text{-}j} + \mathbf{\Gamma}_{k\text{-}i} - \mathbf{\Gamma}_{j\text{-}k}\right] \tag{36}$$

Were " $\underset{\{in3\}}{\approx}$ " indicates the assumption of independence between all three datasets involved.

In the scalar case, Eq. (36) reduces to the equivalent formulation for error variances known from the TC and 3CH method (e.g., Pan et al., 2015; Sjoberg et al., 2021). Thus, the long-known 3CH estimation of error variances from innovation variances between three datasets holds similarly for complete error covariance matrices from innovation covariances under the independent assumption. In fact, the approximation in Eq. (36) requires only the assumption that the dependent contribution of Eq. (26) vanishes. However combining this condition for the error covariance estimates of all three datasets results in the need for each error dependency to be zero.

Under the assumption of independence between all three datasets, their error covariance matrices can also be directly estimated from innovation cross-covariances:

$$\mathbf{C}_{\widetilde{i}} \underset{\{in3\}}{\overset{(29)_{j;l}}{\approx}} \mathbf{\Gamma}_{i\text{-}j;i\text{-}l} \tag{37}$$

And likewise:

$$\mathbf{C}_{\widetilde{i}} \underset{\{in3\}}{\overset{(29)_{l;j}}{\approx}} \mathbf{\Gamma}_{i\text{-}l;i\text{-}j} \tag{38}$$

https://doi.org/10.5194/egusphere-2022-996
Preprint. Discussion started: 10 October 2022

As described in Sect. 3.3.2 on exact cross-covariance statistics, every error covariance from innovation cross-covariances has four equivalent formulations for each pair of other datasets which provide the same result in the exact case, but might differ in the approximative formulation. Equation (37) and Eq. (38) provide two different approximations of each error covariance matrix from innovation cross-covariances based on each pair of other datasets. In the simplified case of scalar statistics, the two different formulations in Eq. (37) and Eq. (38) reduce to the same innovation cross-variance which was previously formulated by e.g. Crow and van den Berg (2010); Zwieback et al. (2012); Pan et al. (2015).

4.1.2 Differences

Equations (36) to (38) provide three different estimates of a error covariance matrix for each pair of other datasets. Using the relation between innovation covariances and -cross-covariances from Sect. 3.2.3 and the symmetric properties of innovation statistics allow a comparison of the three estimates:

$$\mathbf{C}_{\widetilde{i}}\Big|_{(37)} \overset{(37)}{\underset{\{in3\}}{\approx}} \boldsymbol{\Gamma}_{i\text{-}j;i\text{-}l} \overset{(24),(36)}{=} \mathbf{C}_{\widetilde{i}}\Big|_{(36)} + \frac{1}{2}\mathbf{Y}_{i\text{-}j;i\text{-}l} \tag{39}$$

$$\mathbf{C}_{\widetilde{i}}\Big|_{(38)} \overset{(38)}{\underset{\{in3\}}{\approx}} \boldsymbol{\Gamma}_{i\text{-}l;i\text{-}j} \overset{(24),(36)}{=} \mathbf{C}_{\widetilde{i}}\Big|_{(36)} - \frac{1}{2}\mathbf{Y}_{i\text{-}j;i\text{-}l} \tag{40}$$

The three independent estimates of a error covariance matrix from the same pair of other datasets differ only by their innovation asymmetry. Thus, differences between the estimates from Eq. (36) to Eq. (38) provide no additional information about symmetric error statistics.

While the estimation from innovation covariances remains symmetric by definition, the estimates of error covariances from innovation cross-covariances may become asymmetric. This asymmetry can be eliminated using the innovation asymmetry matrix which is also equivalent to averaging both formulations of error covariances from innovation cross-covariances:

$$\mathbf{C}_{\widetilde{i}} \overset{(36)}{\underset{\{in3\}}{\approx}} \frac{1}{2}\Big[\boldsymbol{\Gamma}_{i\text{-}j} + \boldsymbol{\Gamma}_{l\text{-}i} - \boldsymbol{\Gamma}_{j\text{-}l}\Big] \overset{(39)}{=} \boldsymbol{\Gamma}_{i\text{-}j;i\text{-}l} - \frac{1}{2}\mathbf{Y}_{i\text{-}j;i\text{-}l} \overset{(40)}{=} \boldsymbol{\Gamma}_{i\text{-}l;i\text{-}j} + \frac{1}{2}\mathbf{Y}_{i\text{-}j;i\text{-}l} \tag{41}$$

All three estimates become equivalent if the innovation cross-covariances and thus, error cross-covariances are symmetric ($\rightarrow \mathbf{X}_{\widetilde{i};\widetilde{j}} = \frac{1}{2}\mathbf{D}_{\widetilde{i};\widetilde{j}} = \mathbf{X}_{\widetilde{j};\widetilde{i}}\ , \forall\ i,j$). This is also the case for scalar statistics, were the equivalence between scalar error variance estimates from innovation variances and -cross-variances was previously shown by Pan et al. (2015).

4.1.3 Uncertainties of approximation

The independent assumption introduces the following absolute uncertainties $\Delta\mathbf{C}_{\widetilde{i}}$ of the three different estimates for each dataset i:

$$\Delta\mathbf{C}_{\widetilde{i}}\Big|_{(36)} := \mathbf{C}_{\widetilde{i}}\Big|_{\text{true}} - \mathbf{C}_{\widetilde{i}}\Big|_{(36)} \overset{(26),(36)}{=} \frac{1}{2}\Big[\Delta\mathbf{D}_{\widetilde{i};\widetilde{j}} + \Delta\mathbf{D}_{\widetilde{i};\widetilde{k}} - \Delta\mathbf{D}_{\widetilde{j};\widetilde{k}}\Big] \tag{42}$$

$$\Delta\mathbf{C}_{\widetilde{i}}\Big|_{(37)} := \mathbf{C}_{\widetilde{i}}\Big|_{\text{true}} - \mathbf{C}_{\widetilde{i}}\Big|_{(37)} \overset{(29),(37)}{=} \Delta\mathbf{X}_{\widetilde{j};\widetilde{i}} + \Delta\mathbf{X}_{\widetilde{i};\widetilde{k}} - \Delta\mathbf{X}_{\widetilde{j};\widetilde{k}} \tag{43}$$

$$\Delta\mathbf{C}_{\widetilde{i}}\Big|_{(38)} := \mathbf{C}_{\widetilde{i}}\Big|_{\text{true}} - \mathbf{C}_{\widetilde{i}}\Big|_{(38)} \overset{(29),(38)}{=} \Delta\mathbf{X}_{\widetilde{i};\widetilde{j}} + \Delta\mathbf{X}_{\widetilde{k};\widetilde{i}} - \Delta\mathbf{X}_{\widetilde{k};\widetilde{j}} \tag{44}$$

https://doi.org/10.5194/egusphere-2022-996
Preprint. Discussion started: 10 October 2022

were $\Delta\mathbf{D}_{\widetilde{i};\widetilde{j}}$ and $\Delta\mathbf{X}_{\widetilde{i};\widetilde{j}}$ are the uncertainties of the estimated error dependencies and cross-covariances, respectively.

The absolute uncertainty of the estimates depends similarly on the (neglected) error cross-covariances or -dependencies between the three datasets. While the error dependencies to the two other datasets contribute positively, the dependency between the two others is subtracted. If these dependencies cancel out ($\Delta\mathbf{D}_{\widetilde{i};\widetilde{j}} + \Delta\mathbf{D}_{\widetilde{i};\widetilde{k}} = \Delta\mathbf{D}_{\widetilde{j};\widetilde{k}}$), the estimate of one dataset might be exact even if all three dependencies are non-zero. However, two exact estimates can only be achieved if one (e.g. $\Delta\mathbf{D}_{\widetilde{i};\widetilde{j}} = 0 \wedge \Delta\mathbf{D}_{\widetilde{i};\widetilde{k}} = \Delta\mathbf{D}_{\widetilde{j};\widetilde{k}}$) or all three dependencies are zero.

Estimated error covariances might even contain negative values, if error dependencies are large compared to the true error covariance of a dataset. If the true error covariances differ significantly between highly correlated datasets, the neglected error dependency between two datasets might become much larger than the smaller error covariance, e.g. $\Delta\mathbf{D}_{\widetilde{k};\widetilde{i}} - \Delta\mathbf{D}_{\widetilde{j};\widetilde{k}} \approx 0$, $\frac{1}{2}\Delta\mathbf{D}_{\widetilde{i};\widetilde{j}} > \mathbf{C}_{\widetilde{i}}\Big|_{\text{true}}$. Thus, the estimated error covariance matrices might not be positive definite if the independent assumption between three datasets is not fulfilled. This phenomena was also described and demonstrated by Sjoberg et al. (2021) for scalar problems. However, the generalization to covariances matrices is expected to increase the occurrence of negative values were correlations between two entries of the state are low, thus relative differences and sampling errors become large.

4.2 Approximation for multiple datasets

While independence between all datasets is required to estimate the error covariances of three datasets ($I = 3$), the use of more than three datasets ($I > 3$) enables the additional estimation of some error dependencies or cross-covariances (compare Sect. 2). Although this potential of cross-statistic estimation was previously indicated by Gruber et al. (2016); Vogelzang and Stoffelen (2021) for scalar problems, a generalized formulation exploiting its full potential by minimizing the number of assumptions is yet missing.

As described in Sect. 2, $D_I > 0$ gives the number of error cross-statistics which can potentially be estimated in addition to all error covariances for $I > 3$ datasets. Consequently, for each additional dataset $i > 3$, its cross-statistics to one prior dataset ref$(i) < i$ is needed to be assumed in order to close the problem. This prior dataset ref(i) is denoted as "reference dataset" of dataset i. In the following, the approximative estimation of error covariances and -cross-statistics (cross-covariances or dependencies) under the "partly independent assumption" $\mathbf{D}_{\widetilde{i};\widetilde{\text{ref}(i)}} = 0$ is formulated for all additional dataset ($\forall\ i > 3$). This estimation procedure of error statistics of additional dataset based on their reference datasets is denoted as "sequential estimation" in contrast to the "triangular estimation" from an independent triple of datasets presented in Sect. 4.1.

4.2.1 Error covariance estimates

As in the estimation for $I = 3$ datasets, the error covariances of the first three datasets can be estimated from innovation covariances or -cross-covariances using Eq. (36), Eq. (37) or Eq. (38). This triple of the first three datasets which are assumed to be pairwise independent is denoted as "basic triangle".

Based on this, each additional error covariance can directly be calculated w.r.t. its reference dataset ref$(i) < i$:

$$\mathbf{C}_{\widetilde{i}} \overset{(25)_{i,\text{ref}(i)}}{\underset{\{inI\}}{\approx}} \mathbf{\Gamma}_{i\text{-ref}(i)} - \mathbf{C}_{\widetilde{\text{ref}(i)}} \tag{45}$$

https://doi.org/10.5194/egusphere-2022-996
Preprint. Discussion started: 10 October 2022

were " $\underset{\{inI\}}{\approx}$ " indicates the assumption of independence to the reference dataset.

Similarly, each additional error covariance can be estimated from two innovation cross-covariances w.r.t its reference dataset $\mathrm{ref}(i)$ and any other dataset j:

$$\mathbf{C}_{\widetilde{i}} \underset{\{inI\}}{\overset{(31)_{i,\mathrm{ref}(i),j}}{\approx}} \mathbf{\Gamma}_{i-\mathrm{ref}(i);i-j} + \mathbf{\Gamma}_{\mathrm{ref}(i)-i;\mathrm{ref}(i)-j} - \mathbf{C}_{\widetilde{\mathrm{ref}(i)}} \tag{46}$$

From the equivalence of innovation statistics in Eq. (32) is follows that the two formulations of error covariances in Eq. (45) and Eq. (46), respectively, are equivalent and produce exactly the same estimates even if the underlying assumptions are not perfectly fulfilled.

4.2.2 Error cross-covariance and -dependency estimates

Once the error covariances are estimated, the remaining innovation covariances can be used to calculate the error dependencies to all other prior datasets $j \neq \mathrm{ref}(i), j < i$:

$$\mathbf{D}_{\widetilde{i};\widetilde{j}} \overset{(27)_{ij}}{=} \mathbf{C}_{\widetilde{i}} + \mathbf{C}_{\widetilde{j}} - \mathbf{\Gamma}_{i-j} \tag{47}$$

In contrast to innovation covariances, the asymmetric formulation of innovation cross-covariances allows for an estimation of remaining error cross-covariances including their asymmetries. The error cross-covariance to each other prior dataset $j \neq \mathrm{ref}(i), j < i$ can be estimated sequentially using again the reference dataset $\mathrm{ref}(i)$:

$$\mathbf{X}_{\widetilde{i};\widetilde{j}} \underset{\{inI\}}{\overset{(33)_{\mathrm{ref}(i),i,j}}{\approx}} \mathbf{\Gamma}_{\mathrm{ref}(i)-i;\mathrm{ref}(i)-j} - \mathbf{C}_{\widetilde{\mathrm{ref}(i)}} + \mathbf{X}_{\widetilde{\mathrm{ref}(i)};\widetilde{j}} \tag{48}$$

Based on this, the symmetric error dependencies can be estimated from its definition in Eq. (13). The equivalence between the formulations of error dependencies from innovation covariances and cross-covaiances was shown in Eq. (34).

Note that the error cross-covariances $\mathbf{X}_{\widetilde{j};\widetilde{i}}$ and dependencies $\mathbf{D}_{\widetilde{j};\widetilde{i}}$ of each subsequent datasets $j > i$ to dataset j result directly from their symmetric properties in Eq. (10) and Eq. (14), respectively.

4.2.3 Uncertainties in approximation

The absolute uncertainty $\Delta\mathbf{C}_{\widetilde{i}}$ of an additional error covariance estimate of any dataset $3 < i < I$ is formulated recursively w.r.t. its reference dataset $\mathrm{ref}(i)$:

$$\Delta\mathbf{C}_{\widetilde{i}}\Big|_{(45)} := \mathbf{C}_{\widetilde{i}}\Big|_{\mathrm{true}} - \mathbf{C}_{\widetilde{i}}\Big|_{(45)} \overset{(25),(45)}{=} \Delta\mathbf{D}_{\widetilde{i};\widetilde{\mathrm{ref}(i)}} - \Delta\mathbf{C}_{\widetilde{\mathrm{ref}(i)}} \tag{49}$$

$$\Delta\mathbf{C}_{\widetilde{i}}\Big|_{(46)} := \mathbf{C}_{\widetilde{i}}\Big|_{\mathrm{true}} - \mathbf{C}_{\widetilde{i}}\Big|_{(46)} \overset{(31),(46)}{=} \Delta\mathbf{D}_{\widetilde{i};\widetilde{\mathrm{ref}(i)}} - \Delta\mathbf{C}_{\widetilde{\mathrm{ref}(i)}} \tag{50}$$

The two estimates of error covariances from innovation covariances in Eq. (49) and from cross-covariances in Eq. (50) are equivalent, and the uncertainty of the latter is independent of the selection of the third dataset j in the innovation cross-covariances (compare Eq. (46)). Thus, absolute uncertainties of estimations from innovation covariances and -cross-covariances differ only in the uncertainties w.r.t. the basic triangle given in Eq. (42) to Eq. (44).

https://doi.org/10.5194/egusphere-2022-996
Preprint. Discussion started: 10 October 2022

With this, a series of reference datasets $\{m_f\} = m_1, \ldots, m_F$, with m_F beeing the reference of i, and the m_{f-1} reference of m_F and so on, with $m_{f-1} < m_f < i\ , \forall f$ and $m_1 = j \leq 3$ are defined from the target dataset to the basic triangle. Then, the absolute uncertainty $\Delta\mathbf{C}_{\widetilde{i}}$ of each error covariance estimate is:

$$\begin{aligned}\Delta\mathbf{C}_{\widetilde{i}} &\overset{(49)}{=} \Delta\mathbf{D}_{\widetilde{i};\widetilde{m_F}} - \Delta\mathbf{C}_{\widetilde{m_F}} = \Delta\mathbf{D}_{\widetilde{i};\widetilde{m_F}} - \Delta\mathbf{D}_{\widetilde{m_F};\widetilde{m_{F-1}}} + \Delta\mathbf{C}_{\widetilde{m_{F-1}}} = \ldots \\ &\overset{(42)}{=} \Delta\mathbf{D}_{\widetilde{i};\widetilde{m_F}} + \sum_{f=F-1}^{1}\left[(-1)^{F-f}\cdot\Delta\mathbf{D}_{\widetilde{m_{f+1}};\widetilde{m_f}}\right] + (-1)^F\cdot\frac{1}{2}\left[\Delta\mathbf{D}_{\widetilde{j};\widetilde{k}} + \Delta\mathbf{D}_{\widetilde{j};\widetilde{l}} - \Delta\mathbf{D}_{\widetilde{k};\widetilde{l}}\right] \end{aligned} \tag{51}$$

Were $k, l \leq 3$ are the other two datasets in the basic triangle.

According to Eq. (51), uncertainties in the estimations of additional error covariances result from the partly independent assumption of the additional datasets in the series of reference datasets and the independent assumption in the basic triangle. Due to the changing sign between the intermediate dependencies as well as within the basic triangle, the individual uncertainties may cancel out. Thus, absolute uncertainties do not necessarily increase with more intermediate reference datasets.

Although Eq. (47) is exact, the dependency estimate of each additional pair of datasets $(i;j)$ is influenced by uncertainties in the estimations of the related error covariances:

$$\Delta\mathbf{D}_{\widetilde{i};\widetilde{j}} := \mathbf{D}_{\widetilde{i};\widetilde{j}}\Big|_{\text{true}} - \mathbf{D}_{\widetilde{i};\widetilde{j}}\Big|_{(47)} \overset{(27),(47)}{=} \Delta\mathbf{C}_{\widetilde{i}} + \Delta\mathbf{C}_{\widetilde{j}} \tag{52}$$

Were the uncertainties of the two error covariances are given in Eq. (51).

And the absolute uncertainties of estimates of additional error cross-covariances based on innovation cross-covariances can be determined recursively using Eq. (51):

$$\Delta\mathbf{X}_{\widetilde{i};\widetilde{j}} := \mathbf{X}_{\widetilde{i};\widetilde{j}}\Big|_{\text{true}} - \mathbf{X}_{\widetilde{i};\widetilde{j}}\Big|_{(48)} \overset{(33)_{\text{ref}(i),i,j},(48)}{=} \Delta\mathbf{X}_{\widetilde{\text{ref}(i)};\widetilde{j}} + \Delta\mathbf{X}_{\widetilde{i};\widetilde{\text{ref}(i)}} - \Delta\mathbf{C}_{\widetilde{\text{ref}(i)}} \tag{53}$$

In contrast to error covariances, the uncertainties of error cross-covariances sum up in the two series of reference datasets. However, this sum is subtracted by the two sums of uncertainties in error covariances of these datasets, whose elements may cancel partly (not shown).

4.2.4 Comparison to approximation from three datasets

It can be shown that the sequential formulation of an error covariance from its reference dataset is consistent with the triangular formulation from three independent datasets in Sect. 4.1 in the basic triangle. Given the triangular estimate of one error covariance $\mathbf{C}_{\widetilde{i}}\big|_{\triangleleft}$ from Eq. (36), the error covariances $\mathbf{C}_{\widetilde{j}}\big|_{\triangleleft}$ of the other two datasets in the basic triangle are equal to their sequential formulation $\mathbf{C}_{\widetilde{j}}\big|_{\vdash}$ from Eq. (45) with reference dataset $\text{ref}(j) = i$:

$$\mathbf{C}_{\widetilde{j}}\Big|_{\vdash} \underset{\{inI\}}{\overset{(45)_{j,i}}{\approx}} \mathbf{\Gamma}_{i\text{-}j} - \mathbf{C}_{\widetilde{i}}\Big|_{\triangleleft} \underset{\{in3\}}{\overset{(36)_i}{\approx}} \mathbf{\Gamma}_{j\text{-}i} - \frac{1}{2}\left[\mathbf{\Gamma}_{i\text{-}j} + \mathbf{\Gamma}_{k\text{-}i} - \mathbf{\Gamma}_{j\text{-}k}\right] = \frac{1}{2}\left[\mathbf{\Gamma}_{i\text{-}j} + \mathbf{\Gamma}_{j\text{-}k} - \mathbf{\Gamma}_{i\text{-}k}\right] \overset{(36)_j}{=} \mathbf{C}_{\widetilde{j}}\Big|_{\triangleleft} \tag{54}$$

Thus, only one error covariance needs to be calculated with Eq. (36) while all other can be estimated from Eq. (45). Note that although even if only $\mathbf{C}_{\widetilde{i}}$ is calculated from the fully independent formulation in the basic triangle, the independent assumption between all three pairs of datasets in the basic triangle remains.

Instead of using the sequential estimation for additional datasets $i > 3$, the error covariances could also be estimated by defining another independent triangle $(i;j;k)$, with $k = \mathrm{ref}(j)$, $j = \mathrm{ref}(i)$. Because the definition of another independent triangle requires an additional independent assumption between $(i;k)$ (i.e. $\mathbf{D}_{\widetilde{i};\widetilde{k}} = 0$), this triangular estimate $\mathbf{C}_{\widetilde{i}}\big|_{\triangleleft}$ from Eq. (36) differs from the sequential estimate $\mathbf{C}_{\widetilde{i}}\big|_{\vdash}$ from Eq. (45) using its reference dataset ($\mathbf{C}_{\widetilde{j}} \rightarrow \mathbf{C}_{\widetilde{i}}$), were their absolute errors compare as follows:

$$\left|\Delta\mathbf{C}_{\widetilde{i}}\big|_{\vdash}\right| - \left|\Delta\mathbf{C}_{\widetilde{i}}\big|_{\triangleleft}\right| \overset{(42),(49)}{=} \left|\Delta\mathbf{D}_{\widetilde{i};\widetilde{j}} - \Delta\mathbf{C}_{\widetilde{j}}\right| - \frac{1}{2}\left|\Delta\mathbf{D}_{\widetilde{i};\widetilde{j}} + \Delta\mathbf{D}_{\widetilde{i};\widetilde{k}} - \Delta\mathbf{D}_{\widetilde{j};\widetilde{k}}\right| \tag{55}$$

The sequential estimation of error covariances of an error covariance becomes favourable if the estimation of the error covariance of its reference dataset is of similar accuracy as the uncertainty in their dependent assumption $\left(\Delta\mathbf{C}_{\widetilde{j}} \rightarrow \Delta\mathbf{D}_{\widetilde{i};\widetilde{j}}\right)$. And the triangular estimation becomes favourable if the accuracy of the additional independent assumption is of the order of the difference between the uncertainties of other two error dependencies $\left(\Delta\mathbf{D}_{\widetilde{i};\widetilde{k}} \rightarrow \Delta\mathbf{D}_{\widetilde{i};\widetilde{j}} - \Delta\mathbf{D}_{\widetilde{j};\widetilde{k}}\right)$; i.e. if the additional independent assumption is of similar accuracy as the other two dependent assumptions.

Note that the absolute uncertainties presented here only account for uncertainties due to the underlying assumptions on error cross-statistics and not due to imperfect innovation statistics occurring e.g. from finite sampling. An discussion of those effects for scalar problems can be found in Sjoberg et al. (2021).

5 Experiments

This section provides an exemplary demonstration of the capabilities to estimate full error covariance matrices of all datasets and some error dependencies. Four collocated datasets ($I = 4$) are generated synthetically on a 1D domain with 25 gridpoints. Each datset consists of 20.000 realizations at each gridpoint which are randomly sampled around the true value of 5.0 . The spatial variation of prescribed error variances and spatial error correlations differ for each dataset. Datasets $(1;2;3)$ span the basic triangle and dataset 1 is the reference of dataset 4 ($\mathrm{ref}(4) = 1$). This allows the estimation of all four error covariances of each dataset and two error dependencies between the datasets (2;4) and (3;4) (compare Sect. 2). The experiments presented in this section are based on the symmetric estimations from innovation covariances derived in Sect. 4 which are summarized in Algorithm A1. All other error dependencies need to be assumed and are set to zero for this experiment (in accordance to the formulation in Sect. 4). Note that the change of error dependencies between the different experiments affects their true statistics.

The plots are structured as follows: Each subplot combines two covariance matrices; one shown in the upper-left part and the other in the lower-right part. The two matrices are separated by a gray bar and shifted off-diagonal so that diagonal variances are right above/below the gray bar, respectively. Statistics that might become negative are shown as absolute quantities in order to show them with the same color-code. In each row, the upper-left parts are matrices which are usually unknown in real applications (as they require the knowledge of the truth) and the lower-right parts are known/estimated matrices. The first row contains the error dependencies and innovation covariances between each dataset pair. Here, gray asterisks in the upper-left subplot indicate that these error dependency matrices are assumed to be zero in the estimation. The second row contains the true and estimated error covariances and dependencies. The third row gives the absolute difference between the true and estimates matrices.

5.1 Uncertainties in additional dependencies

The experiment shown in Fig. 2a contains only true error dependencies between datasets $(2;4)$ and $(3;4)$. This is consistent to the selected estimation setup in which $(1;2;3)$ build the basic triangle and dataset 4 is sequentially estimated calculated w.r.t. dataset 1. Consequently, all error covariances and the two remaining dependencies are estimated accurately in accordance to Eq. (42), Eq. (51), and Eq. (52). The estimation method is able to reproduce true error covariance matrices of all datasets and error dependency matrices between some datasets independent of the complexity of the statistics if the assumptions are sufficiently fulfilled. For comparison, the error covariance matrix of dataset 4 is estimated from an additional independent triangle $(1;2;4)$ in Fig. 2b. The triangular estimation requires the additional independent assumption between datasets $(2;4)$ which is not fulfilled in this experiment. The positive true error dependency has equivalent impact on the estimated error covariance of dataset 4 and its dependencies to datasets 2 and 3. All three matrices are underestimated w.r.t. the true statistics by the half of the neglected error dependency in accordance to Eq. (42), and Eq. (52) applied to the trianlge $(1;2;4)$.

5.2 Uncertainties in basic triangle

The effects of neglected dependencies in the basic triangle is exemplary demonstrated in Fig. 3 were a true positive error dependency appears between datasets $(2;3)$. In accordance to Eq. (42), Eq. (51), and Eq. (52), the neglected dependency in the basic triangle affects all estimated statistics. While the error covariance of dataset 1 is overestimated, all other statistics are underestimated. For the sequential estimation in Fig. 3a, uncertainties in the estimated error dependencies are equal to the neglected error dependency and uncertainties in error covariances are halved (compare Eq. (51)). For the triangular estimation of dataset 4 in Fig. 3b, the effects of the two neglected dependencies between $(2;3)$ and between $(2;4)$ are combined. As shown in Fig. 2b, the error covariance of dataset 4 is underestimated by half the neglected dependency between $(2;4)$. The uncertainties two estimated error dependencies $(2;4)$ and $(3;4)$ are the sum of the uncertainties of the error covariances of the two datasets involved in accordance to Eq. (52).

In this setup, the sequential estimation of the additional dataset (here 4) from its reference dataset (here 1) is more accurate because the neglected dependency in the basic triangle (here $(2;3)$) is small compared to the neglected additional dependency in the triangular estimation (here $(2;4)$), which is in accordance to Eq. (55). This changes in Fig. 4, were the neglected dependency (here $(2;3)$) in the basic triangle is larger than the one additional one in the triangular estimation (here $(2;4)$). Because the sequential estimation is more sensitive to uncertainties in the basic triangle (Fig. 4a), the triangular estimation (Fig. 4b) becomes more accurate. This holds for the error covariance estimate of dataset 4 as well as the two estimated error dependencies $(2;4)$ and $(3;4)$.

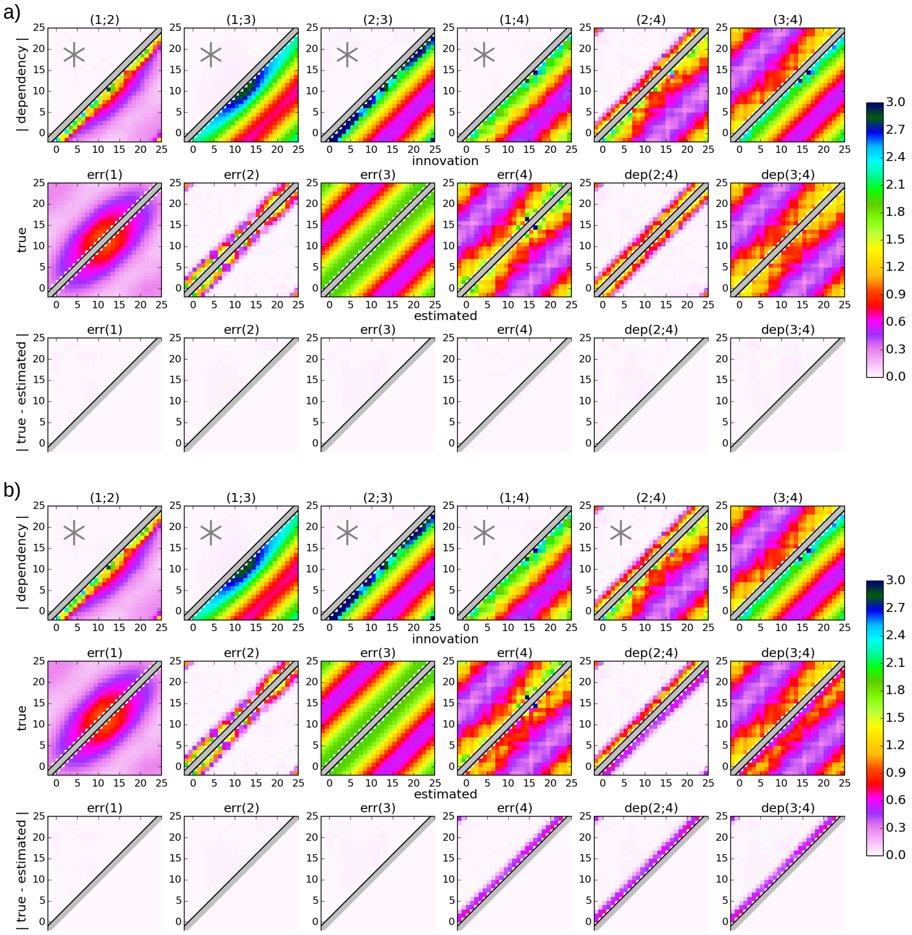

Figure 2. Covariance matrices for 4 datasets ($I = 4$) with true dependencies of datasets (2;4) and (3;4). Datasets (1;2;3) build the basic triangle. Dataset 4 is estimated (a) from its reference dataset 1 ("sequential estimation") and (b) from an additional independent triangle (1;2;4) ("triangular estimation").

Note that the choice of the estimation method does only affect the uncertainty of subsequent estimates which are directly or indirectly referring to the uncertain assumption. In this case, the estimations in the basic triangle are not affected by the estimation method.

6 Conceptual summary

This section provides a summary of the statistical error estimation method proposed in this study focusing on it's technical application. Section 6.1 summarises the general requirements of assumptions including an exemplary visualisation, and Sect. 6.2

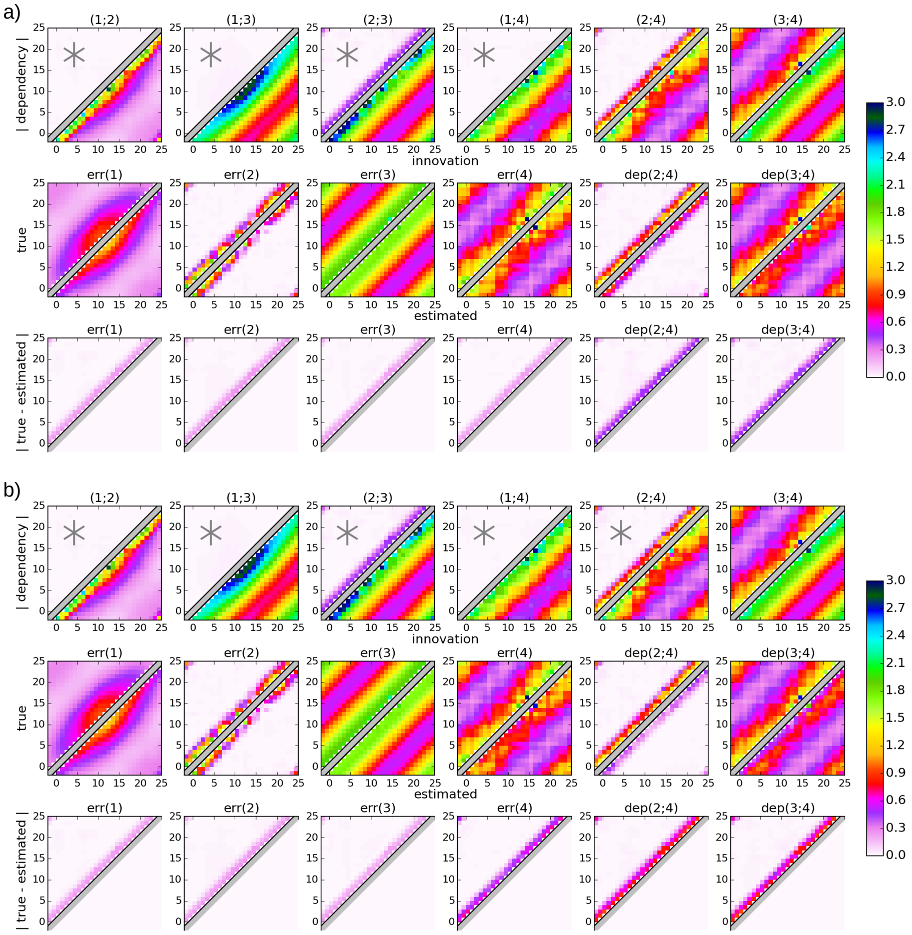

Figure 3. Covariance matrices for 4 datasets ($I = 4$) with true dependencies of datasets (2;3), (2;4) and (3;4). As in Fig. 2, but with an neglected dependency in the basic triangle between datasets (2;3).

formulates rules for an optimal setup of datasets w.r.t. imperfect assumptions. An algorithmic summary the calculation of error statistics from innovation covariances and cross-covariances, respectively, is given in Apx. A.

6.1 Minimal conditions

For error statistics that need to be assumed, their specific formulation may have different forms. The easiest and most common assumption is to set their error correlations and thus the error cross-covariances and dependencies to zero. This assumption used in Sect. 4.1 and 4.2 and is equivalent to the 3CH and TC methods. However, any non-zero error statistics can be defined and used in the general form which is summarized in Apx. A. This also includes assuming error statistics as function of other error statistics including the ones estimated during the calculation. The only restriction is that all assumed error statistics must be fully determined by other error statistics or predefined values without introducing additional degrees of freedom.

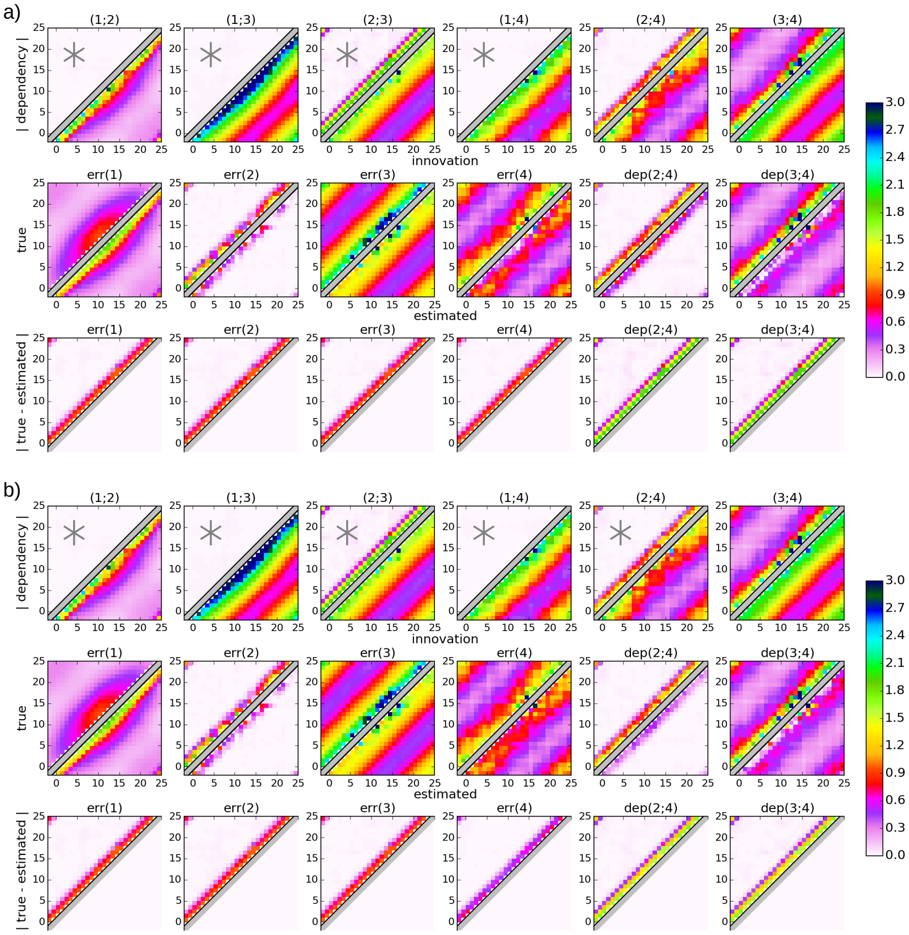

Figure 4. Covariance matrices for 4 datasets ($I = 4$) with true dependencies of datasets (2;3), (2;4) and (3;4). As in Fig. 2, but with an increased dependency between in the basic triangle between datasets (2;3).

In the common case were all error covariances and some error dependencies (or cross-covariances) are estimated, there are two requirements for the setup of datasets: (i) all three error dependencies between one triple of datasets are needed (this triple of independent datasets is called "basic triangle"), and (ii) at least one error dependency of each additional dataset to any prior datasets is needed (this prior dataset is called "reference dataset" of the referring additional datasets). Previously, Vogelzang and Stoffelen (2021) observed that some setups for four and five datasets do not produce a solution of the problem, but without discussing the general requirements. The limited solveability was also found by Gruber et al. (2016) for four datasets, who whoever came up with a too strong requirement that each dataset has to be part of an independent triangle.

An exemplary setup of assumed dependencies for $I = 10$ datasets is visualized in Fig. 5. The dependencies between three datasets (1;2;3) is needed to be assumed ("basic triangle"). Then, one dependency of each additional dataset $i > 3$ to any

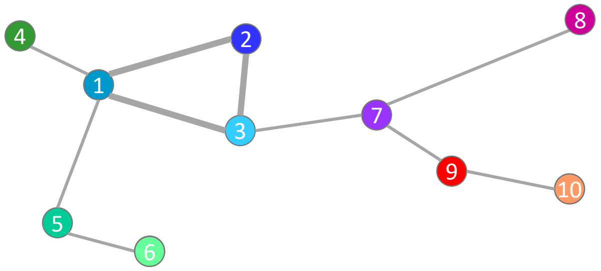

Figure 5. Independence tree: Exemplary visualization of assumed dependencies (gray lines) between 10 datasets (colored dots). The assumed dependencies in the basic triangle (1;2;3) are indicated by thicker lines.

prior dataset j (with $j < i$) is assumed ("sequential estimation"). In general, there is no further restriction on the selection of reference datasets in order to close the estimation problem.

6.2 Optimal setup

In real applications, there might be significant differences in estimated error statistics from different setups as observed e.g. by Vogelzang and Stoffelen (2021) in the scalar case. The relative accuracy of an error covariance estimates is proportional to the ratio between the innovation covariance $\mathbf{\Gamma}_{i\text{-}j}$ and the absolute uncertainty $\Delta\mathbf{D}_{\tilde{i};\tilde{j}}$ of the assumed error dependency, which can be interpreted similar to a signal-to-noise ratio. In other words, the larger the innovation covariance and the better the absolute estimate of the error dependency to the reference dataset, the more accurate is the estimated error covariance. Because uncertainties in error estimate do not necessarily sum up along a branch of the independence tree (compare Sect. 4.2.3, Eq. (51)), a large innovation-to-dependency ratio w.r.t. to the reference is more important than a low number of intermediate reference datasets. In order to achieve optimal estimates, the setup of datasets should be selected according to the expected accuracy of estimated dependencies which minimize the innovation-to-dependency ratio for each datasets:

$$\max_j \left(\frac{\mathbf{\Gamma}_{i\text{-}j}}{\Delta\mathbf{D}_{\tilde{i};\tilde{j}}} \right) \; : j \rightarrow \mathrm{ref}(i) \quad \Longleftrightarrow \; \min_j \left(\Delta\rho_{\tilde{i};\tilde{j}} \right) \; : j \rightarrow \mathrm{ref}(i) \quad , \forall\, i \tag{56}$$

The maximal innovation-to-dependency ratio is equivalent to the minimal uncertainty in normalized error correlations $\Delta\rho_{\tilde{i};\tilde{j}} := \frac{\Delta\mathbf{D}_{\tilde{i};\tilde{j}}}{\sqrt{\mathbf{C}_{\tilde{i}}\,\mathbf{C}_{\tilde{j}}}}$. For example, if the error correlation of one dataset to another is comparably well known, this dataset is best suited as reference dataset. If estimated error dependencies are set to zero, the dataset to which the independent assumption is most certain should be selected as reference dataset. Supposing that distances between datasets indicate their expected degree of independence in the independence tree, the setup visualized in Fig. 5 is not optimal. An example for an improved setup is shown in Fig. 6, which is expected to provide more accurate error estimates.

While uncertainties in the basic triangle are only contribute half, they effect the estimations of error statistics of all datasets (compare Sect. 4.1.3, 4.2.3 and 5.2). This has two implications: Firstly, the basic triangle which is defined as the triple of dataset that has the smallest error correlations produces the smallest overall uncertainty w.r.t. all error estimates. Ideally, the

https://doi.org/10.5194/egusphere-2022-996
Preprint. Discussion started: 10 October 2022

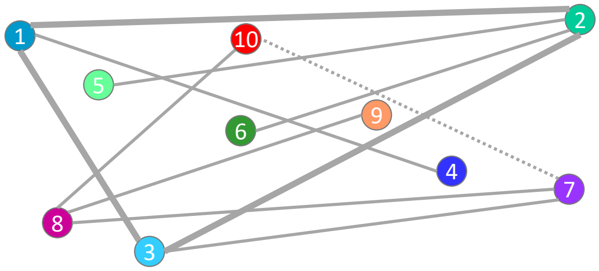

Figure 6. Improved independence tree: As Fig. 5, but with modified setup for more accurate error estimates. Distances between datasets represent the accuracy of assumed dependencies between the error statistics. While locations are the same, the numbers and colors of the datasets has been changed according to the modified setup. An example for an alternative formulation of an additional independent triangle is indicated as dotted line.

basic triangle should be set as a triple of datasets which are highly independent – or at least with reasonably small dependencies among each other.

Secondly, if another independent triangle can be assumed for an additional dataset with similar accuracy as the dependency to its reference dataset, the additional error estimate may be more accurate using the triangular estimation from this additional independent triangle rather than the sequential estimation (compare Sect. 4.2.4 and Sect. 5). The additional independent triangle does not need to be connected to the basic triangle and may also have multiple independence branches, thus acting as additional basic triangle. For example, in the setup shown in Fig. 6, the estimation of dataset 7 is sensitive to the dependency(3;7) to its reference dataset, and less sensitive to dependencies in the basic triangle (1;2;3). If the dependency(7;10) could be assumed with higher accuracy than these dependencies, the error covariances of dataset 7 can alternatively be calculated from the independent triangle (7;8;10) and the independent assumption between (3;7) can be dropped. Thus, multiple independence trees can be defined around multiple separated basic triangles.

Furthermore, it is also possible to average the estimated error statistics of a dataset from multiple independent triangles similar to an application of the N-cornered hat method (e.g., Sjoberg et al., 2021) for an arbitrary subset of datasets. This setup builds an overestimated problem which requires the assumption of more error dependencies than the minimal requirements. However, it might be beneficial if multiple independent triangles containing the same dataset can be estimated with similar accuracy. In this case, potential uncertainties in the assumptions are expected to be reduced by the average over similar accurate estimates. Also an extension to weighted averages of different estimations is possible were the weights reflect the expected accuracy of each estimation formulation w.r.t. the others.

7 Conclusions

Despite the generalized matrix-formulation, the main features of the of the presented approach are (i) its generality defining the flexible setup for any number of datasets according to the specific application, (ii) its optimality w.r.t. a minimal number of assumptions required, and (iii) its suitability to include expected non-zero dependencies between any pair of datasets. In contrast, the scalar N-CH method averages all estimates of each dataset which is equivalent to assuming that the independent assumption between each dataset triple is fulfilled with the same accuracy. However, this is not the case for most applications to geophysical datasets. For example, Rieckh et al. (2021) applied the N-CH method to multiple atmospheric model and observational datasets and discussed neglected levels of independence between different datasets, which are expected to vary significantly. Pan et al. (2015) tried to account for such variations by clustering the datasets into structural groups; which however requires more assumptions than necessary and makes the result highly sensitive to the selected grouping. In contrast, the method presented here provides an optimal and flexible approach to handle multiple datasets with different levels of expected independence. Depending on the specific application, the estimation may be based on the minimal number of assumptions required or a (weighted) average over any number of estimations with similar expected accuracies.

In comparison to a-posteriori methods which statistically estimate optimal error covariances for data assimilation, an a-priori error estimation of collocated datasets has three main advantages: (i) optimal error statistics are calculated analytically without requiring an iterative minimization including multiple executions of the assimilation, (ii) complete covariance matrices provide spatially-resolved fields of error statistics at each collocated location including spatial and cross-species correlations, and (iii) error statistics of all datasets are estimated without selecting one dataset as reference. This enables the consideration of more than two datasets in the assimilation. Given sufficiently estimated error statistics, the final analysis w.r.t. to all datasets will be closer to the truth than any analysis between two datasets only. Thus, the rapidly increasing number of geophysical observations and model forecast enables improved analyses through increasingly overlapping datasets, were optimal error statistics can be calculated for example with the method presented here. Especially the possibility to estimate optimal error cross-covariances between datasets provides important information for data assimilation were the violation of the independent assumption remains a mayor challenge (Tandeo et al., 2020).

However, current data assimilation schemes are not suited for multiple overlapping datasets and cross-errors between datsets are assumed to be neglectable. In contrast, the statistical error estimation method presented in this study is explicitly tailored to multiple datasets which cannot be assumed to be independent. Thus, the estimated error covariances are not consistent with assimilation algorithms assuming (two) independent datasets. If the estimated error dependencies between all assimilated datasets are small, the independent assumption may be regarded as sufficiently fulfilled. The error estimation method then provides optimal error covariances for assimilation and information on the accuracy of the independent assumption. Otherwise, generalized assimilation schemes are need to be developed for a proper use of this additional statistical information in data assimilation. Although increasing their complexity, such generalized assimilation schemes enable fundamental improvements in terms of an optimal analysis from multiple datasets w.r.t. their error covariances and cross-statistics.

Appendix A: Algorithm

The general estimation procedure of error statistics for $I \geq 3$ datsets is summarized in Algorithm A1 and A2. The algorithms require respectively, innovation covariances or cross-covariances between all I datsets (calculated from innovation statistics) and I assumed error dependencies cross-covariances. Based on this, the first error covariance matrix in the basic triangle is calculated. Then, error statistics of the remaining datasets are calculated sequentially in an iterative procedure; introducing a new dataset i with given innovation statistics (covariances or cross-covariances) to dataset $\mathrm{ref}(i)$ for each $i \in [2, I]$ with $\mathrm{ref}(i) < i$. Note that this is equivalent to estimating the independent estimations of all three datasets in basic triangle and sequentially estimate all additional estimates for datsets $i > 3$ (compare Sect. 4.2.4).

Algorithm A1 is formulated for symmetric statistic matrices, were error covariances $\texttt{errcov}(i;:,:)$ of each dataset i and error dependency matrices $\texttt{errdep}(i;j;:;:)$ between each pair $(i;j)$ are estimated from symmetric innovation covariances $\texttt{innocov}(i\text{-}j\ :;:;)$. In Algorithm A2, the error covariance and cross-covariance matrices $\texttt{errcross}(i;j;:;:)$ of each pair $(i;j)$ are estimated from innovation cross-covariances $\texttt{innocross}(i\text{-}j; i\text{-}k;:;:)$ between $(i\text{-}j; i\text{-}k)$. Here, the third dataset k in the innovation cross-covariances can be freely selected and does not affect the accuracy of the estimates (compare Sect. 4.2.3). Each operation applies elementwise to each matrix-element indicated by the last two indices $(:;:)$, were matrices may contain different locations of the same quantity as well as different fields for multiple quantities of any dimension (=multivariate covariances). Transposed matrices w.r.t. the two location indices are indicated by $[\,]^T$.

The equations relate to the general exact formulations which requires some error dependencies or cross-covariances to be given (compare Sect. 3). The explicit calculation of the error cross-statistics (dependencies or cross-covariances) is not needed if only error covariances are of interest. In theory, both algorithms provide the same error estimations (compare Sect.3.2.3). The decision to estimate error statistics from innovation covariances (Algorithm A1) or cross-covariances (Algorithm A2) depends on the availability of innovation statistics and the need for asymmetric error cross-covariances, which can only be estimated with Algorithm A2 (compare Sect. 3.3.1).

Algorithm A1 Iterative calculation of error covariances and dependencies for I datasets from innovation covariances.

```
Require: $\texttt{innocov}(i\text{-ref}(i);:;:)\ \forall\ i \in [2, I]$, $\texttt{innocov}(1\text{-}3;:;:)$
Require: $\texttt{errdep}(i;\mathrm{ref}(i);:;:)\ \forall\ i \in [2, I]$, $\texttt{errdep}(1;3;:;:)$

  $\texttt{errcov}(1;:;:) \leftarrow 0.5 \cdot \Big[\texttt{innocov}(2\text{-}1;:;:) + \texttt{innocov}(1\text{-}3;:;:) - \texttt{innocov}(3\text{-}2;:;:)$
                $+\texttt{errdep}(2;1;:;:) + \texttt{errdep}(1;3;:;:) - \texttt{errdep}(3;2;:;:)\Big]$          ▷ ∼ Eq. (26)
  for $i = 2, I$ do
      $\texttt{errcov}(i;:;:) \leftarrow \texttt{innocov}(i\text{-ref}(i);:;:) + \texttt{errdep}(i;\mathrm{ref}(i);:;:) - \texttt{errcov}(\mathrm{ref}(i);:;:)$          ▷ ∼ Eq. (25)
      for $j = 1, i-1$ do
          if $j \neq \mathrm{ref}(i)$ then
              $\texttt{errdep}(i;j;:;:) \leftarrow \texttt{errcov}(i;:;:) + \texttt{errcov}(j;:;:) - \texttt{innocov}(i\text{-}j;:;:)$          ▷ ∼ Eq. (27)
          end if
          $\texttt{errdep}(j;i;:;:) \leftarrow \texttt{errdep}(i;j;:;:)$          ▷ ∼ Eq. (14)
      end for
  end for
```

Algorithm A2 Iterative calculation of error covariances and cross-covariances for I datasets from innovation cross-covariances.

```
Require: $\texttt{innocross}(i\text{-ref}(i); i\text{-}j; :; :), \texttt{innocross}(\text{ref}(i)\text{-}i; \text{ref}(i)\text{-}j; :; :)\ \forall\ i \in [2, I], j \neq \text{ref}(i), j \neq i\ , \texttt{innocross}(1\text{-}2; 1\text{-}3; :; :)$
Require: $\texttt{errcross}(i; \text{ref}(i); :; :)\ \forall\ i \in [2, I]$, $\texttt{errcross}(1; 3; :; :)$

  for $i = 2, I$ do
      $\texttt{errcross}(\text{ref}(i); i; :; :) \leftarrow \texttt{errcross}(i; \text{ref}(i); :; :)^T$                ▷ ~ Eq. (10)
  end for
  $\texttt{errcov}(1; :; :) \leftarrow \texttt{innocross}(1\text{-}2; 1\text{-}3; :; :) + \texttt{errcross}(1; 3; :; :) + \texttt{errcross}(2; 1; :; :) - \texttt{errcross}(2; 3; :; :)$   ▷ ~ Eq. (29)
  for $i = 2, I$ do
      $\texttt{errcov}(i; :; :) \leftarrow \texttt{innocross}(i\text{-ref}(i); i\text{-}j; :; :) + \texttt{innocross}(\text{ref}(i)\text{-}i; \text{ref}(i)\text{-}j; :; :) - \texttt{errcov}(\text{ref}(i); :; :)$
                       $+\texttt{errcross}(i; \text{ref}(i); :; :) + \texttt{errcross}(\text{ref}(i); i; :; :)$   ▷ ~ Eq. (31)
      for $j = 1, i - 1$ do
          if $j \neq \text{ref}(i)$ then
              $\texttt{errcross}(i; j; :; :) \leftarrow \texttt{innocross}(\text{ref}(i)\text{-}i; \text{ref}(i)\text{-}j; :; :) - \texttt{errcov}(\text{ref}(i); :; :)$
                          $+\texttt{errcross}(\text{ref}(i); j; :; :) + \texttt{errcross}(i; \text{ref}(i); :; :)$   ▷ ~ Eq. (33)
              $\texttt{errcross}(j; i; :; :) \leftarrow \texttt{errcross}(i; j; :; :)^T$   ▷ ~ Eq. (10)
          end if
      end for
  end for
```

Author contributions. AV developed the approach, derived the theory, performed the experiments, and wrote the manuscript. RM supervised the work, and revised the manuscript.

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

References

Anthes, R. and Rieckh, T.: Estimating observation and model error variances using multiple data sets, Atmospheric Measurement Techniques, 11, 4239–4260, https://doi.org/10.5194/amt-11-4239-2018, 2018.

Crow, W. T. and van den Berg, M. J.: An improved approach for estimating observation and model error parameters in soil moisture data assimilation, Water Resources Research, 46, https://doi.org/https://doi.org/10.1029/2010WR009402, 2010.

Crow, W. T. and Yilmaz, M. T.: The Auto-Tuned Land Data Assimilation System (ATLAS), Water Resources Research, 50, 371–385, https://doi.org/https://doi.org/10.1002/2013WR014550, 2014.

Daley, R.: The Effect of Serially Correlated Observation and Model Error on Atmospheric Data Assimilation, Monthly Weather Review, 120, 164 – 177, https://doi.org/10.1175/1520-0493(1992)120<0164:TEOSCO>2.0.CO;2, 1992a.

Daley, R.: The Lagged Innovation Covariance: A Performance Diagnostic for Atmospheric Data Assimilation, Monthly Weather Review, 120, 178 – 196, https://doi.org/10.1175/1520-0493(1992)120<0178:TLICAP>2.0.CO;2, 1992b.

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

Gray, J. and Allan, D.: A Method for Estimating the Frequency Stability of an Individual Oscillator, in: 28th Annual Symposium on Frequency Control, pp. 243–246, https://doi.org/10.1109/FREQ.1974.200027, 1974.

Grubbs, F. E.: On Estimating Precision of Measuring Instruments and Product Variability, Journal of the American Statistical Association, 43, 243–264, https://doi.org/10.1080/01621459.1948.10483261, 1948.

Gruber, A., Su, C.-H., Crow, W. T., Zwieback, S., Dorigo, W. A., and Wagner, W.: Estimating error cross-correlations in soil moisture data sets using extended collocation analysis, Journal of Geophysical Research: Atmospheres, 121, 1208–1219, https://doi.org/https://doi.org/10.1002/2015JD024027, 2016.

Kren, A. C. and Anthes, R. A.: Estimating Error Variances of a Microwave Sensor and Dropsondes aboard the Global Hawk in Hurricanes Using the Three-Cornered Hat Method, Journal of Atmospheric and Oceanic Technology, 38, 197 – 208, https://doi.org/10.1175/JTECH-D-20-0044.1, 2021.

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

McColl, K. A., Vogelzang, J., Konings, A. G., Entekhabi, D., Piles, M., and Stoffelen, A.: Extended triple collocation: Estimating errors and correlation coefficients with respect to an unknown target, Geophysical Research Letters, 41, 6229–6236, https://doi.org/https://doi.org/10.1002/2014GL061322, 2014.

Ménard, R.: Error covariance estimation methods based on analysis residuals: theoretical foundation and convergence properties derived from simplified observation networks, Quarterly Journal of the Royal Meteorological Society, 142, 257–273, https://doi.org/https://doi.org/10.1002/qj.2650, 2016.

Ménard, R. and Deshaies-Jacques, M.: Evaluation of Analysis by Cross-Validation. Part I: Using Verification Metrics, Atmosphere, 9, https://doi.org/10.3390/atmos9030086, 2018.

Mitchell, H. L. and Houtekamer, P. L.: An Adaptive Ensemble Kalman Filter, Monthly Weather Review, 128, 416 – 433, https://doi.org/10.1175/1520-0493(2000)128<0416:AAEKF>2.0.CO;2, 2000.

Nielsen, J. K., Gleisner, H., Syndergaard, S., and Lauritsen, K. B.: Estimation of refractivity uncertainties and vertical error correlations in collocated radio occultations, radiosondes and model forecasts, Atmospheric Measurement Techniques Discussions, 2022, 1–28, https://doi.org/10.5194/amt-2022-121, 2022.

Pan, M., Fisher, C. K., Chaney, N. W., Zhan, W., Crow, W. T., Aires, F., Entekhabi, D., and Wood, E. F.: Triple collocation: Beyond three estimates and separation of structural/non-structural errors, Remote Sensing of Environment, 171, 299–310, https://doi.org/https://doi.org/10.1016/j.rse.2015.10.028, 2015.

Rieckh, T., Sjoberg, J. P., and Anthes, R. A.: The Three-Cornered Hat Method for Estimating Error Variances of Three or More Atmospheric Datasets. Part II: Evaluating Radio Occultation and Radiosonde Observations, Global Model Forecasts, and Reanalyses, Journal of Atmospheric and Oceanic Technology, 38, 1777 – 1796, https://doi.org/10.1175/JTECH-D-20-0209.1, 2021.

Scipal, K., Holmes, T., de Jeu, R., Naeimi, V., and Wagner, W.: A possible solution for the problem of estimating the error structure of global soil moisture data sets, Geophysical Research Letters, 35, https://doi.org/https://doi.org/10.1029/2008GL035599, 2008.

Sjoberg, J. P., Anthes, R. A., and Rieckh, T.: The Three-Cornered Hat Method for Estimating Error Variances of Three or More Atmospheric Datasets. Part I: Overview and Evaluation, Journal of Atmospheric and Oceanic Technology, 38, 555 – 572, https://doi.org/10.1175/JTECH-D-19-0217.1, 2021.

Stoffelen, A.: Toward the true near-surface wind speed: Error modeling and calibration using triple collocation, Journal of Geophysical Research: Oceans, 103, 7755–7766, https://doi.org/https://doi.org/10.1029/97JC03180, 1998.

Su, C.-H., Ryu, D., Crow, W. T., and Western, A. W.: Beyond triple collocation: Applications to soil moisture monitoring, Journal of Geophysical Research: Atmospheres, 119, 6419–6439, https://doi.org/https://doi.org/10.1002/2013JD021043, 2014.

Tandeo, P., Ailliot, P., Bocquet, M., Carrassi, A., Miyoshi, T., Pulido, M., and Zhen, Y.: A Review of Innovation-Based Methods to Jointly Estimate Model and Observation Error Covariance Matrices in Ensemble Data Assimilation, Monthly Weather Review, 148, 3973 – 3994, https://doi.org/10.1175/MWR-D-19-0240.1, 2020.

Tangborn, A., Ménard, R., and Ortland, D.: Bias correction and random error characterization for the assimilation of high-resolution Doppler imager line-of-sight velocity measurements, Journal of Geophysical Research: Atmospheres, 107, ACL 5–1–ACL 5–15, https://doi.org/https://doi.org/10.1029/2001JD000397, 2002.

Todling, R., Semane, N., Anthes, R., and Healy, S.: The relationship between two methods for estimating uncertainties in data assimilation, Quarterly Journal of the Royal Meteorological Society, https://doi.org/https://doi.org/10.1002/qj.4343, 2022.

Vogelzang, J. and Stoffelen, A.: Quadruple Collocation Analysis of In-Situ, Scatterometer, and NWP Winds, Journal of Geophysical Research: Oceans, 126, e2021JC017 189, https://doi.org/https://doi.org/10.1029/2021JC017189, 2021.

Voshtani, S., Ménard, R., Walker, T. W., and Hakami, A.: Assimilation of GOSAT Methane in the Hemispheric CMAQ; Part I: Design of the Assimilation System, Remote Sensing, 14, https://doi.org/10.3390/rs14020371, 2022.

Xu, X. and Zou, X.: Global 3D Features of Error Variances of GPS Radio Occultation and Radiosonde Observations, Remote Sensing, 13, https://doi.org/10.3390/rs13010001, 2021.

Zwieback, S., Scipal, K., Dorigo, W., and Wagner, W.: Structural and statistical properties of the collocation technique for error characterization, Nonlinear Processes in Geophysics, 19, 69–80, https://doi.org/10.5194/npg-19-69-2012, 2012.