Received: 22 November 2021 | Revised: 25 May 2022 | Accepted: 6 June 2022 | Published on: 31 July 2022

DOI: 10.1002/qj.4339

Quarterly Journal of the Royal Meteorological Society RMetS

RESEARCH ARTICLE

New impact diagnostics for cross-validation of different observation types

Olaf Stiller

Data Assimilation Section, Deutscher Wetterdienst (DWD), Franfurt am Main, 63067, Germany

Correspondence
Olaf Stiller, Data Assimilation Section, Deutscher Wetterdienst (DWD), 63067 Franfurt am Main, Germany.
Email: Olaf.Stiller@dwd.de

Abstract

New cross-validation diagnostics have been derived by further partitioning well-established impact diagnostics. They are related to consistency relations, the most prominent of which indicates whether the first-guess departures of a given observation type pull the model state into the direction of the verifying data (when processed with the ensemble estimated model error covariances). Alternatively, this can be regarded as cross-validation between model error covariance estimates from the ensemble (which are used in the data assimilation system) and estimates diagnosed directly from the observations. A statistical cross-validation tool has been developed that includes an indicator of statistical significance as well as a normalization that makes the statistical comparison largely independent from the total number of data and the closeness of their collocation. We also present a version of these diagnostics related to single-observation experiments that exploits the same consistency relations but is easier to compute. Diagnostics computed within the Deutscher Wetterdienst's localized ensemble transform Kalman filter (LETKF) are presented for various kinds of bins. Results from well-established in-situ measurements are taken as a benchmark for more complex observations. Good agreement is found for radio-occultation bending angle measurements, whereas atmospheric motion vectors are generally also beneficial but substantially less optimal than the corresponding in-situ measurements. This is consistent with reported atmospheric motion vector height assignment problems. To illustrate its potential, a recent example is given where the method allowed identifying bias problems of a subgroup of aircraft measurements. Another diagnostic relationship compares the information content of the analysis increments with a theoretical optimum. From this, the information content of the LETKF increments is found to be considerably lower than those of the deterministic hybrid ensemble–variational system, which is consistent with the LETKF's limitation to the comparably low-dimensional ensemble space for finding the optimal analysis.

KEYWORDS

analysis information content, consistency relations, ensemble covariance, ensemble Kalman filter, observation impact, optimality condition

1 INTRODUCTION

The quality of numerical weather prediction (NWP) strongly depends on the proper use of observational data for improving the forecast model's initial conditions through the data assimilation (DA) process. Modern DA methods ingest a great variety of data types that correspond to measurements on very different scales and which are generally distributed inhomogeneously in space and time.

Assessing the impact that these data have on the NWP process is important, but rigorous tests are numerically very expensive. Fortunately, methods have been developed that offer a reasonably cheap approximate impact measure. Here, the pioneering work goes back to Langland and Baker (2004), who started from a verification function (or forecast metric) and presented a method how a proxy for the contribution of the different observation types to this metric can be computed at very low additional cost (called forecast sensitivity to observation impact [FSOI]). Though the original method required the adjoint of the forecast model (which is available in a four-dimensional variational system), Liu and Kalnay (2008) showed how the same type of verification metric can be evaluated using the ensemble of an ensemble Kalman filter. Later, Kalnay *et al.* (2012) showed how such computations of an ensemble-based FSOI (EFSOI) can be achieved in an even simpler (and numerically cheaper) way. An approach for computing an FSOI for a hybrid system was presented by Buehner *et al.* (2018).

Although so far, in most of the work in the field, forecast impact is measured via verification against an analysis state (i.e., verification in state space), Sommer and Weissmann (2016) and Necker *et al.* (2018) used an observation-based measure instead—which had also been proposed by Todling (2013)—which they tested within the ensemble system of the German Weather Service (Deutscher Wetterdienst, DWD). Also, Cardinali (2018) used an observation-based forecast metric with the adjoint-based system of the European Centre for Medium-Range Weather Forecasts. Though the spatial and temporal coverage of observations is less homogeneous, verifying against them has the advantage that it avoids the problem of self-verification—resulting from correlations of forecast errors with those of the state used for verification; for example, see Kotsuki *et al.* (2019). Using observations as a proxy for the truth (rather than an analysis state) particularly permits the verification at shorter forecast lead times. Even the verification at the analysis time (forecast lead-time zero) is possible.

There are two main motivations for computing the impact of observations in an NWP system. First, particularly in times of limited resources, knowing the impact of the different parts of the observing system is crucial for planning and decision-making with respect to the ongoing efforts and developments concerning the generation, distribution, and processing of observational data. To this end, the focus is usually on the relative importance (or ranking) of the different observation types, and some studies have been made to show that, for the ranking, FSOI-type statistics may give results similar to what is obtained from denial experiments.

The second motivation stems from the fact that exploiting observations in an NWP system is not always trivial and all the underlying mathematical concepts rely on various theoretical assumptions, many of which are violated by most real-world observations. Making optimal usage of a given observation type requires great efforts and error mitigation measures, which depend on both the respective observations and the DA system employed. To this end, (E)FSOI-type statistics have been used to design quality control methods (Hotta *et al.*, 2017a; Chen and Kalnay, 2019), to tune the error-covariance matrix (Hotta *et al.*, 2017b), and, generally, to facilitate the usage of new data types in a DA system (Lien *et al.*, 2018).

The work presented here revisits the established (E)FSOI statistics to develop diagnostic tools for identifying observation types (or groups of observations) that are processed suboptimally by the respective DA system. For this, the established (E)FSOI diagnostics are split into two parts, each of which is related to a different aspect of the data-processing system and for which different consistency relationships are derived. These consistency relationships provide reference values for these diagnostics and also allow some interpretation outside the FSOI context. Further, we provide a normalization that renders results largely independent from the total number of observations and the closeness of their collocation and, also, an indicator of statistical significance (which reflects both the magnitude and the number of the different contributions).

The aim of the diagnostics discussed herein is not just to assess whether the impact is beneficial but also to give some idea to what extent the processing of an observation type is consistent with the expectations and assumptions that are the theoretical basis of the DA system. Though we do not expect these theoretical assumptions to be met precisely, the strategy of this work is to first gain some experience of the kind of correspondence that can be expected between such idealized theoretical values and the statistics gathered from real-world data. Working with the local ensemble transform Kalman filter (LETKF)—see Hunt *et al.* (2007)—that is operational at the DWD, we start by looking at well-established observation types that are known to have a clearly beneficial impact in our DA system. (This also gives some interesting insights into the functioning and limitations of the DA system.) Starting with in-situ measurements from radiosondes (TEMPs) and aircraft, we show differences with the more complex GPS radio occultation (GPSRO) measurements; and we also demonstrate, in the example of wind measurements from atmospheric motion vectors (AMVs), how some less-optimal behavior of these more indirect measurements can be identified by comparing their statistics with those from TEMPs and aircraft. Also, as another example for the method's potential, we present some recent results where the cross-validation method allowed identifying bias problems with a type of aircraft temperature measurements in the DWD's DA system.

Compared with more traditional cross-validation applications—for example, as developed by Ménard and Deshaies-Jacques (2018a) and Ménard and Deshaies-Jacques (2018b)—the main difference of the new diagnostics is the partitioning of the corresponding verification function, which is called "mean-square-error costfunction" by Ménard and Deshaies-Jacques (2018a). In the following, we make use of the FSOI-type partitioning to allow us to identify contributions from different observation types from a single analysis (i.e., without performing denial experiments), plus a further partitioning producing impact-related diagnostics that can be interpreted with respect to consistency relations. However, in contrast to some previously introduced residual-based consistency diagnostics (e.g., Hollingsworth and Lönnberg, 1986; Desroziers *et al.*, 2005), our aim is not to estimate observation errors and their covariance but to diagnose the performance that an observation type has in the DA process and whether this is consistent with other observations. For this, the non-diagonal elements of the background error covariance in observation space play an essential role; indeed, our new diagnostics can also be used for validating the background error covariances employed. This, however, is restricted to the cross-covariances between statistically independent observations for which these observations give us a direct estimator.

This article is structured as follows. The mathematical foundations for the new diagnostics and their interpretation are given in Section 2. This part is largely independent of whether FSOI is computed for an ensemble or variational system and whether observations or an analysis state are used for verification. A more detailed description of the corresponding diagnostic strategy that has been developed for our LETKF system is given in Section 3, and the results related to some major observation types are presented in Section 4. Section 5 discusses these results in the context of our diagnostic strategy, and some of the avenues that this will involve in the future are described in Section 5.1.

2 | MATHEMATICAL FOUNDATIONS

2.1 | FSOI-type statistics

FSOI-type statistics start from a verification function J that quantifies the impact which the observations $\boldsymbol{y}^o$ assimilated at the analysis time t_0 have on the forecast at time $t_0 + t$. Here, $\boldsymbol{y}^o$ is a vector whose dimension is the number of all assimilated observations. Using further the vector $\boldsymbol{y}^{\mathbf{v}}$ for the data used for verification (at time $t_0 + t$), while $\boldsymbol{y}^{\mathbf{v}|a}$ and $\boldsymbol{y}^{\mathbf{v}|b}$ are the corresponding model equivalents (computed from model forecasts initialized at analysis time t_0 with the analysis $\boldsymbol{x}^a$ and the first guess $\boldsymbol{x}^b$, respectively) we define the verification function

$$J \overset{\text{def!}}{=} \frac{1}{2}(\|\boldsymbol{y}^{\mathbf{v}} - \boldsymbol{y}^{\mathbf{v}|a}\|^2 - \|\boldsymbol{y}^{\mathbf{v}} - \boldsymbol{y}^{\mathbf{v}|b}\|^2). \tag{1}$$

This exhibits negative values (indicating beneficial observation impact) if the forecast starting from the analysis ($\boldsymbol{y}^{\mathbf{v}|a}$) fits the verification data $\boldsymbol{y}^{\mathbf{v}}$ better than that starting from the background state ($\boldsymbol{y}^{\mathbf{v}|b}$). Though, for the results presented herein, observations are used for the verification (i.e., $\boldsymbol{y}^{\mathbf{v}}$ are observations made at verification time $t_0 + t$), the mathematical relations derived herein are more general. To this end, $\boldsymbol{y}^{\mathbf{v}}$ can be any vector-type quantity that serves for verification (e.g., like an analysis state at time $t_0 + t$). The metric $\| \ldots \|$ is then defined via the scalar product

$$\|\boldsymbol{y}^{\mathbf{v}} - \boldsymbol{y}^{\mathbf{v}|a}\|^2 = (\boldsymbol{y}^{\mathbf{v}} - \boldsymbol{y}^{\mathbf{v}|a})^{\mathrm{T}} C^{-1} (\boldsymbol{y}^{\mathbf{v}} - \boldsymbol{y}^{\mathbf{v}|a}),$$

where the symmetric and positive definite matrix C must be adequate for the chosen verification space.

The standard FSOI procedures partition this verification function into a sum of the form

$$J = \sum_{\alpha \in \{\text{obs}\}} J_\alpha \tag{2}$$

whose terms are attributed as the contribution of the respective observation α (i.e., the α component of $\boldsymbol{y}^o$) to the verification function J. For the work presented here, the components J_α are further partitioned into two different components that can be related to different consistency relations and which yield information about different aspects of the observations and how they are used in the DA system. As shown in the following, we can write

$$J_\alpha = -\frac{1}{2}[2J_\alpha^b - J_\alpha^{ab}] \tag{3}$$

with

$$J^b_\alpha = (\boldsymbol{y}^{\mathbf{v}} - \boldsymbol{y}^{\mathbf{v}|b})^{\mathrm{T}} \boldsymbol{C}^{-1} \widehat{\boldsymbol{P}}^{a}\{\mathbf{v}, o\} \boldsymbol{R}^{-1} \boldsymbol{\Pi}_\alpha (\boldsymbol{y}^o - \boldsymbol{y}^b) \tag{4a}$$

$$J^{ab}_\alpha = (\boldsymbol{y}^{\mathbf{v}|a} - \boldsymbol{y}^{\mathbf{v}|b})^{\mathrm{T}} \boldsymbol{C}^{-1} \widehat{\boldsymbol{P}}^{a}\{\mathbf{v}, o\} \boldsymbol{R}^{-1} \boldsymbol{\Pi}_\alpha (\boldsymbol{y}^o - \boldsymbol{y}^b) \tag{4b}$$

where $\boldsymbol{R}$ is the employed observation error covariance matrix and $\boldsymbol{\Pi}_\alpha$ is a projection operator that sets all components of $\boldsymbol{y}^o$ apart from α to zero.

Here, we have introduced the matrix

$$\widehat{\boldsymbol{P}}^{a}\{\mathbf{v}, o\} \equiv \boldsymbol{H}^{\mathbf{v}} \boldsymbol{M}_t \mathbf{P}^a \mathbf{H}^{\mathrm{T}} \tag{5}$$

where $\mathbf{H}$ and $\mathbf{P}^a$ are the linearized observation operator and the employed error-covariance matrix of the analysis state, $\boldsymbol{M}_t$ is the time evolution operator (between analysis time t_0 and verification time $t_0 + t$), and $\boldsymbol{H}^{\mathbf{v}}$ is the linearized version of the operator that computes the model equivalents $\boldsymbol{y}^{\mathbf{v}|a}$ and $\boldsymbol{y}^{\mathbf{v}|b}$ from the corresponding forecast model states at verification time. When verifying with observations (which is the case for all the applications discussed in this article), $\boldsymbol{H}^{\mathbf{v}}$ is the corresponding observation operator linearized about the background state; for verification against analysis, $\boldsymbol{H}^{\mathbf{v}}$ is the identity operator (or possibly some interpolation operator if the verifying analysis is on a different grid than the forecast model that is tested).

Note that if the different factors on the right (i.e., the operators $\boldsymbol{M}_t$, $\boldsymbol{H}^{\mathbf{v}}$, and $\mathbf{H}^{\mathrm{T}}$ as well as the covariance matrix $\mathbf{P}^a$) were all fully correct (which in particular implies that the corresponding linear approximations must be fully valid), the quantity $\widehat{\boldsymbol{P}}^{a}\{\mathbf{v}, o\}$ would be the error covariance matrix (or more precisely the cross-covariance matrix) between the verification space vector $\boldsymbol{y}^{\mathbf{v}|a}$ (which is valid at verification time $t_0 + t$) and the observation-space vector $\boldsymbol{y}^a$ (the analysis in observation space valid at t_0); that is,

$$\widehat{\boldsymbol{P}}^{a}\{\mathbf{v}, o\} = \mathrm{cov}(\boldsymbol{e}\{\boldsymbol{y}^{\mathbf{v}|a}\}, \boldsymbol{e}\{\boldsymbol{y}^a\})$$

where $\boldsymbol{e}\{:\}$ denotes the error, or difference from the truth, of the respective quantity.

In the following, to derive Equations 4a and 4b, we use the notation

$$\boldsymbol{e}^a = \boldsymbol{y}^{\mathbf{v}|a} - \boldsymbol{y}^{\mathbf{v}}$$
$$\boldsymbol{e}^b = \boldsymbol{y}^{\mathbf{v}|b} - \boldsymbol{y}^{\mathbf{v}}$$

to write

$$\begin{aligned} J &= \frac{1}{2}[(\boldsymbol{e}^a)^{\mathrm{T}} \boldsymbol{C}^{-1} \boldsymbol{e}^a - (\boldsymbol{e}^b)^{\mathrm{T}} \boldsymbol{C} \boldsymbol{e}^b] \\ &= \frac{1}{2}(\boldsymbol{e}^a + \boldsymbol{e}^b)^{\mathrm{T}} \boldsymbol{C}^{-1} (\boldsymbol{e}^b - \boldsymbol{e}^a) \\ &= \frac{1}{2}(\boldsymbol{e}^a + \boldsymbol{e}^b)^{\mathrm{T}} \boldsymbol{C}^{-1} (\boldsymbol{y}^{\mathbf{v}|a} - \boldsymbol{y}^{\mathbf{v}|b}) \\ &= \frac{1}{2}(\boldsymbol{e}^a + \boldsymbol{e}^b)^{\mathrm{T}} \boldsymbol{C}^{-1} \boldsymbol{H}^{\mathbf{v}} \boldsymbol{M}_t \mathbf{K} (\boldsymbol{y}^o - \boldsymbol{y}^b) \end{aligned} \tag{6}$$

where $\mathbf{K}$ is the Kalman gain matrix and $\boldsymbol{y}^b$ the model equivalent of the observation vector $\boldsymbol{y}^o$ corresponding to the first-guess state $\boldsymbol{x}^b$ (i.e., to a short-term forecast). In the last step,

$$\boldsymbol{y}^{\mathbf{v}|a} - \boldsymbol{y}^{\mathbf{v}|b} = \boldsymbol{H}^{\mathbf{v}} \boldsymbol{M}_t \mathbf{K} (\boldsymbol{y}^o - \boldsymbol{y}^b) \tag{7}$$

was used. Rewriting Equation (6) further by introducing

$$\boldsymbol{e}^a + \boldsymbol{e}^b = 2(\boldsymbol{y}^{\mathbf{v}|b} - \boldsymbol{y}^{\mathbf{v}}) + (\boldsymbol{y}^{\mathbf{v}|a} - \boldsymbol{y}^{\mathbf{v}|b}) \tag{8}$$

in the first set of parentheses on the right-hand side, writing the Kalman gain matrix in the form

$$\mathbf{K} = \mathbf{P}^a \mathbf{H}^{\mathrm{T}} \boldsymbol{R}^{-1} \tag{9}$$

or equivalently

$$\boldsymbol{H}^{\mathbf{v}} \boldsymbol{M}_t \mathbf{K} = \widehat{\boldsymbol{P}}^{a}\{\mathbf{v}, o\} \boldsymbol{R}^{-1}$$

and using $\boldsymbol{y}^o - \boldsymbol{y}^b = \sum_\alpha \boldsymbol{\Pi}_\alpha (\boldsymbol{y}^o - \boldsymbol{y}^b)$, the verification function J can be decomposed as given by Equation (2) with Equations 3,4a, and 4b. Note that this result is consistent with what was found by Kalnay *et al.* (2012) and which has been strongly exploited in the literature.

2.2 | The role of the different components of J_α

Owing to the statistical nature of observation and background errors, J_α and its components introduced in Equation (3) are statistical quantities. Therefore, to get meaningful results, one needs to take the average over (or the sum over) a sufficient number of observations, in which case the results are related to the statistical expectation values. In the following, statistical expectation values will be indicated by the use of angle brackets $\langle : \rangle$.

2.2.1 | The cross-validation diagnostic

We would initially like to note that the first component J^b_α of J_α can be independently related to a different type of verification function, which we define as

$$J^b = (\boldsymbol{y}^{\mathbf{v}} - \boldsymbol{y}^{\mathbf{v}|b})^{\mathrm{T}} \boldsymbol{C}^{-1} (\boldsymbol{y}^{\mathbf{v}|a} - \boldsymbol{y}^{\mathbf{v}|b}) \tag{10}$$

and which can be related to J by writing

$$J = -J^b + \frac{1}{2}\|\boldsymbol{y}^{\mathbf{v}|a} - \boldsymbol{y}^{\mathbf{v}|b}\|^2. \tag{11}$$

This can be directly obtained by substituting Equation (8) into the third line of Equation (6). Following the analysis that led to Equation (4a)—that is, introducing Equation (7) with Equation (9) and the definition Equation (5) into Equation (10)—one finds that

$$J^b = \sum_\alpha J^b_\alpha;$$

that is, J^b_α is the α contribution to J^b.

From Equation (11), a beneficial impact from the analysis (i.e., $J < 0$) is only possible if $J^b > 0$. Examining Equation (10), the reason for this is evident, as from its definition, if J^b is negative, the analysis increments $\boldsymbol{y}^{\mathbf{v}|a} - \boldsymbol{y}^{\mathbf{v}|b}$ have the opposite sign to the first-guess departures of $\boldsymbol{y}^{\mathbf{v}}$, which means that the analysis pulls the model state in the opposite direction to the verifying data. We therefore take it as a fundamental necessary condition for an observation type to possibly have a positive impact that its contribution J^b_α to the verification function J^b must (on average) be positive; that is,

$$\langle J^b_\alpha \rangle > 0. \tag{12}$$

Note that this is mainly a condition for the first-guess departures of the observations to which the component J^b_α is most sensitive.

For practical applications, a more sensitive condition for identifying suboptimalities is obtained by comparing J^b_α with a reference value for the magnitude that this diagnostic would assume if basic assumptions made in the derivations of our DA systems were fully fulfilled. To derive such a reference value we rewrite Equation (4a) using the trace function. Given that $\mathrm{Tr}[\mathbf{a}^{\mathrm{T}}\mathbf{b}] = \mathrm{Tr}[\mathbf{b}\mathbf{a}^{\mathrm{T}}]$ (for any vectors or matrices $\mathbf{a}$ and $\mathbf{b}$ that have the same dimension) we can write

$$\langle J^b_\alpha \rangle = \mathrm{Tr}[\boldsymbol{C}^{-1}\widehat{\boldsymbol{P}}^a\{\mathbf{v},o\}\boldsymbol{R}^{-1}\boldsymbol{\Pi}_\alpha\langle(\boldsymbol{y}^o - \boldsymbol{y}^b)(\boldsymbol{y}^{\mathbf{v}} - \boldsymbol{y}^{\mathbf{v}|b})^{\mathrm{T}}\rangle] \tag{13}$$

where the expectation value on the right-hand side can be considered as an estimator for the error cross-covariance between the background error of $\boldsymbol{y}^b$ at analysis time t_0 with that of $\boldsymbol{y}^{\mathbf{v}|b}$ at verification time $t_0 + t$. Using $\epsilon^{\mathbf{v}}$ and ϵ^o for the errors of the verifying data $\boldsymbol{y}^{\mathbf{v}}$ and the assimilated observations $\boldsymbol{y}^o$, respectively, and the superscript "tr" for the true values of the respective quantities, we write for this expectation value

$$\begin{aligned}
&\langle(\boldsymbol{y}^o - \boldsymbol{y}^b)(\boldsymbol{y}^{\mathbf{v}} - \boldsymbol{y}^{\mathbf{v}|b})^{\mathrm{T}}\rangle \\
&= \langle[\epsilon^o - (\boldsymbol{y}^b - \boldsymbol{y}^{\mathrm{tr}})][\epsilon^{\mathbf{v}} - (\boldsymbol{y}^{\mathbf{v}|b} - \boldsymbol{y}^{\mathbf{v}|\mathrm{tr}})]^{\mathrm{T}}\rangle \\
&= \langle(\boldsymbol{y}^b - \boldsymbol{y}^{\mathrm{tr}})(\boldsymbol{y}^{\mathbf{v}|b} - \boldsymbol{y}^{\mathbf{v}|\mathrm{tr}})^{\mathrm{T}}\rangle \\
&= \mathrm{cov}(\boldsymbol{e}\{\boldsymbol{y}^b\}, \boldsymbol{e}\{\boldsymbol{y}^{\mathbf{v}|b}\})
\end{aligned} \tag{14}$$

where it was assumed that all errors $\epsilon^{\mathbf{v}}$ of the verifying data $\boldsymbol{y}^{\mathbf{v}}$ and ϵ^o of the assimilated observations $\boldsymbol{y}^o$ are bias free, uncorrelated with forecast errors, and mutually uncorrelated.

A reference value for $\boldsymbol{J}^b_\alpha$ is now obtained by replacing this observation estimate for the background error covariance by the corresponding quantity from the DA and NWP system. In strict analogy to $\widehat{\boldsymbol{P}}^a\{\mathbf{v},o\}$ in Equation (5), we define

$$\widehat{\boldsymbol{P}}^b\{\mathbf{v},o\} \equiv \boldsymbol{H}^{\mathbf{v}}\boldsymbol{M}_t\mathbf{P}^b\mathbf{H}^{\mathrm{T}}, \tag{15}$$

which would yield

$$\widehat{\boldsymbol{P}}^b\{\mathbf{v},o\} = \mathrm{cov}(\boldsymbol{e}\{\boldsymbol{y}^{\mathbf{v}|b}\}, \boldsymbol{e}\{\boldsymbol{y}^b\})$$

if the respective linear operators and covariance matrices used by the NWP system were fully valid. With this we define the estimator

$$\langle J^b_\alpha \rangle_{\mathrm{estim}} = \mathrm{Tr}[\boldsymbol{C}^{-1}\widehat{\boldsymbol{P}}^a\{\mathbf{v},o\}\boldsymbol{R}^{-1}\boldsymbol{\Pi}_\alpha\widehat{\boldsymbol{P}}^b\{o,\mathbf{v}\}] \tag{16}$$

for $\boldsymbol{J}^b_\alpha$, where $\widehat{\boldsymbol{P}}^b\{o,\mathbf{v}\}$ is the transpose of $\widehat{\boldsymbol{P}}^b\{\mathbf{v},o\}$ defined in Equation (15).

2.2.2 | The optimal information content

The second component $\boldsymbol{J}^{ab}_\alpha$ of $\boldsymbol{J}_\alpha$ is proportional to the size of the analysis increments, which raises the question about what the optimal size of these increments is. Just looking at Equation (3) one might be tempted to aim for a very small (or even negative) value of $\langle \boldsymbol{J}^{ab}_\alpha \rangle$. This, however, neglects the nonlinear nature of the verification function J and that if the analysis state $\boldsymbol{x}^a$ was optimal for initializing the forecast with respect to the verification function J then the following *optimality condition*[1] would be fulfilled:

$$\langle J^b_\alpha \rangle = \langle J^{ab}_\alpha \rangle. \tag{17}$$

One way to see this is to consider the initial conditions

$$\boldsymbol{x}^{\hat{a}} = \boldsymbol{x}^a + \delta\lambda^{(\mathbf{v})}_\alpha \mathbf{K}\boldsymbol{\Pi}_\alpha(\boldsymbol{y}^o - \boldsymbol{y}^b). \tag{18}$$

[1] In this article we are referring to this equation as the *optimality condition*. However, we would like to point out that, as far as we can see, this is only a necessary condition for that the analysis state $\boldsymbol{x}^a$ (or equivalently $\boldsymbol{M}_t\boldsymbol{x}^a$) could be optimal in minimizing the verification function J.

with

$$\delta\lambda_\alpha^{(\mathbf{v})} = \frac{\langle J_\alpha^b - J_\alpha^{ab}\rangle}{\langle \|\boldsymbol{H}^{\mathbf{v}}\boldsymbol{M}_t\mathbf{K}\boldsymbol{\Pi}_\alpha(\boldsymbol{y}^o - \boldsymbol{y}^b)\|^2\rangle}, \tag{19}$$

which coincide with $\boldsymbol{x}^a$ if Equation (17) is fulfilled (in which case $\delta\lambda_\alpha^{(\mathbf{v})} = 0$). Using the linear approximation (which is central to the Kalman filter), the corresponding forecast (in verification space) takes the form

$$\boldsymbol{y}^{\mathbf{v}|\hat{a}} = \boldsymbol{y}^{\mathbf{v}|a} + \delta\lambda_\alpha^{(\mathbf{v})}\boldsymbol{H}^{\mathbf{v}}\boldsymbol{M}_t\mathbf{K}\boldsymbol{\Pi}_\alpha(\boldsymbol{y}^o - \boldsymbol{y}^b), \tag{20}$$

which for $\delta\lambda_\alpha^{(\mathbf{v})} \neq 0$ fits better with the verifying data than $\boldsymbol{y}^{\mathbf{v}|a}$, and then one has (see Appendix A.1 for details):

$$\begin{aligned}\langle \|\boldsymbol{y}^{\mathbf{v}} - \boldsymbol{y}^{\mathbf{v}|\hat{a}}\|^2\rangle &= \langle \|\boldsymbol{y}^{\mathbf{v}} - \boldsymbol{y}^{\mathbf{v}|a}\|^2\rangle - (\delta\lambda_\alpha^{(\mathbf{v})})^2 \\ &\quad \times \langle \|\boldsymbol{H}^{\mathbf{v}}\boldsymbol{M}_t\mathbf{K}\boldsymbol{\Pi}_\alpha(\boldsymbol{y}^o - \boldsymbol{y}^b)\|^2\rangle \\ &< \langle \|\boldsymbol{y}^{\mathbf{v}} - \boldsymbol{y}^{\mathbf{v}|a}\|^2\rangle.\end{aligned} \tag{21}$$

This shows that Equation (17) (or equivalently $\delta\lambda_\alpha^{(\mathbf{v})} = 0$) is a necessary condition for $\boldsymbol{x}^a$ (or equivalently $\boldsymbol{M}_t\boldsymbol{x}^a$) being a global minimum of the verification function J, since otherwise—that is, if Equation (17) is not fulfilled—the initial condition $\boldsymbol{x}^{\hat{a}}$ leads to a smaller value of J. A slightly different proof for this is also given in Appendix A.2, which shows that Equation (17) is also a necessary condition for a *local* minimum at $\boldsymbol{x} = \boldsymbol{x}^a$ (a local minimum requires that any derivative of J with respect to the initial conditions is zero at $\boldsymbol{x} = \boldsymbol{x}^a$).

In Appendix A.3, we further show more specifically that the optimality condition, Equation (17), holds if

$$\begin{aligned}&\widehat{\boldsymbol{P}}^b\{\mathbf{v}, o\}[\hat{\mathbf{P}}^b + \boldsymbol{R}]^{-1}\langle(\boldsymbol{y}^o - \boldsymbol{y}^b)(\boldsymbol{y}^o - \boldsymbol{y}^b)^{\mathrm{T}}\rangle\boldsymbol{\Pi}_\alpha^{\mathrm{T}} \\ &= \langle(\boldsymbol{y}^{\mathbf{v}} - \boldsymbol{y}^{\mathbf{v}|b})(\boldsymbol{y}^o - \boldsymbol{y}^b)^{\mathrm{T}}\rangle\boldsymbol{\Pi}_\alpha^{\mathrm{T}}\end{aligned} \tag{22}$$

is fulfilled (where $\hat{\mathbf{P}}^b \equiv \mathbf{H}\mathbf{P}^b\mathbf{H}^{\mathrm{T}}$). If the mathematical assumptions on which the DA system is based were fully fulfilled, one could write

$$[\hat{\mathbf{P}}^b + \boldsymbol{R}] = \langle(\boldsymbol{y}^o - \boldsymbol{y}^b)(\boldsymbol{y}^o - \boldsymbol{y}^b)^{\mathrm{T}}\rangle \tag{23}$$

for the error covariances of the assimilated observations, so that Equation (22) takes the form

$$\widehat{\boldsymbol{P}}^b\{\mathbf{v}, o\}\boldsymbol{\Pi}_\alpha^{\mathrm{T}} = \langle(\boldsymbol{y}^{\mathbf{v}} - \boldsymbol{y}^{\mathbf{v}|b})(\boldsymbol{y}^o - \boldsymbol{y}^b)^{\mathrm{T}}\rangle\boldsymbol{\Pi}_\alpha^{\mathrm{T}}. \tag{24}$$

Equation (24) can be regarded as a consistency relation for the background error covariance between verification and assimilation space. If it is fulfilled, the covariance used in the DA system $\widehat{\boldsymbol{P}}^b\{\mathbf{v}, o\}$, defined in Equation (15), equals what is diagnosed by the observations, according to Equation (14). Equation (22) can be regarded as a generalization of this consistency relation, which is relevant in situations where Equation (23) is clearly violated. This particularly includes situations where observation errors are inflated, which in practice is often done to compensate for possible suboptimal features (or a suboptimal processing) of some observation types. As computing the left-hand side of Equation (22) is highly non-trivial, in the following we will also consider the corresponding relation for single-observation experiments (for which the vector $\boldsymbol{y}^o$ has only a single component "y_α^o") which include effects from possible error inflation for the observation y_α^o. More precisely, we use

$$\widehat{\boldsymbol{P}}^b\{\mathbf{v}, o\}\boldsymbol{\Pi}_\alpha^{\mathrm{T}}Q_\alpha = \langle(\boldsymbol{y}^{\mathbf{v}} - \boldsymbol{y}^{\mathbf{v}|b})(\boldsymbol{y}^o - \boldsymbol{y}^b)^{\mathrm{T}}\rangle\boldsymbol{\Pi}_\alpha^{\mathrm{T}} \tag{25}$$

with

$$Q_\alpha = \frac{\langle(y_\alpha^o - y_\alpha^b)^2\rangle}{(\hat{P}_{\alpha\alpha}^b + R_{\alpha\alpha})} \tag{26}$$

as another generalization of Equation (24), which in situations where the errors of $\boldsymbol{y}_\alpha^o$ are inflated is more closely related to the optimality condition, Equation (17), than the relation in Equation (24).

3 | FSOI-BASED DIAGNOSTICS FOR AN ENSEMBLE DA SYSTEM

3.1 | The new diagnostics

Applying the analysis from the last section further to the DWD LETKF gave rise to various statistical diagnostics (as shown, for example, in Figure 2), and the principal aim of this section is to make the reader familiar with the different curves in such graphs. For computing these EFSOI-type statistics we follow Sommer and Weissmann (2016) to use observations for verification (i.e., $\boldsymbol{y}^{\mathbf{v}}$ are observations made at verification time $t_0 + t$) with the metric $\boldsymbol{C}$ being the error covariance matrix of these observations. The main challenge for computing the verification function J and its components from Equation 4a and 4b is to estimate the analysis error cross-covariances $\widehat{\boldsymbol{P}}^a\{\mathbf{v}, o\}$, and to this end we use the respective estimates $\widehat{P}^a_{\mathrm{en}[\nu,\alpha]}$ from the LETKF, which, as explained in Appendix B, is the respective covariance $\widetilde{P^a}_{\mathrm{en}[\nu,\alpha]}$ (between y_α^a and $y_\nu^{\mathbf{v}|a}$) from the ensemble multiplied by the localization function $\eta^t(\nu, \alpha)$; that is,

$$\widehat{P}^a_{\mathrm{en[v,\alpha]}} = \widetilde{P^a}_{\mathrm{en}[\nu,\alpha]} * \eta^t(\nu, \alpha). \tag{27}$$

It should be noted that the LETKF actually uses a different type of localization (called R-localization), where, instead of reducing the background covariance to zero for large distances (called B-localization), the observation error grows exponentially as a function of distance between the observations and the region for which the analysis is performed. Since the weight observations are given in the analysis are largely determined by the ratio of the background and observation error variances employed, R-localization generally has a similar effect to B-localization. Equation (27) is best characterized as B-localization in observation space, which is the way localization is usually performed for EFSOI applications—for example, see Houtekamer and Zhang (2016) and references therein for a discussion of different localization types.

As indicated in Equation (B3), the localization function employed in this work is just a superposition of two Gaspari–Cohn functions (Gaspari and Cohn, 1999), where the vertical (l_z) and horizontal (l_h) localization length scales are kept at constant values, with $l_h = 300\,\text{km}$ and the dimensionless value $l_z = 0.3$ for the vertical localization in logarithmic pressure coordinates. This value for l_z differs from the corresponding length scale used in our LETKF, which varies linearly as a function of logarithmic pressure from $l_z = 0.3$ at the surface to $l_z = 0.8$ at the model top. Though for higher levels this difference could lead to an underestimation regarding the full EFSOI (J_α), the interpretation of the components J^b_α and J^{ab}_α in terms of consistency relations remains valid as the same localization is applied to all quantities (including reference values) consistently. Note that in most studies the EFSOI localization differs from the localization employed in the ensemble Kalman filter, not only in terms of localization type (e.g., R- and B-localization) but also in that the FSOI localization includes the model state at verification time (which generally differs from the analysis time; for example, see Gasperoni and Wang (2015) and references therein for a discussion of the consequences).

Using Equation (27) and restricting to the case where observation error covariance matrices are diagonal (with elements $R_{\nu\nu}$ for the verification data and $R_{\alpha\alpha}$ for the assimilated data) one obtains for the verification function components Equations 4a and 4b

$$J^b_\alpha = \sum_\nu \hat{P}^a_{\text{en}[\nu,\alpha]} \frac{(y^{\mathbf{v}}_\nu - y^{\mathbf{v}|b}_\nu)(y^o_\alpha - y^b_\alpha)}{R_{\nu\nu} R_{\alpha\alpha}} \tag{28a}$$

$$J^{ab}_\alpha = \sum_\nu \hat{P}^a_{\text{en}[\nu,\alpha]} \frac{(y^{\mathbf{v}|a}_\nu - y^{\mathbf{v}|b}_\nu)(y^o_\alpha - y^b_\alpha)}{R_{\nu\nu} R_{\alpha\alpha}} \tag{28b}$$

where the sum is over all verification observations $y^{\mathbf{v}}_\nu$. In this case, the corresponding reference value, defined in Equation (16), is given by

$$\langle J^b_\alpha \rangle_{\text{estim}} = \sum_\nu \hat{P}^a_{\text{en}[\nu,\alpha]} \frac{\hat{P}^b_{\text{en}[\nu,\alpha]}}{R_{\nu\nu} R_{\alpha\alpha}}. \tag{29}$$

3.1.1 | Single-observation diagnostics

It should be noted that another extremely useful set of diagnostics that does not require knowledge of the analysis error covariance $\hat{P}^a_{\text{en}[\nu,\alpha]}$ can be easily calculated at the same low cost. These are the corresponding statistics related to single-observation experiments. Though the standard FSOI considered so far is a proxy for the impact that an observation y^o_α has when assimilated simultaneously with all the other observations—and which can be obtained from the last line of Equation (6) by setting all first-guess departures, apart from that for y^o_α, to zero—single-observation statistics yield the impact of y^o_α that would be obtained if no other observation was assimilated. They are obtained by replacing in Equations 28a and 28b the analysis error covariance $\hat{P}^a_{\text{en}[\nu,\alpha]}$ and the analysis increments $(y^{\mathbf{v}|a}_\nu - y^{\mathbf{v}|b}_\nu)$ by their single-observation counterparts:

$$\hat{P}^{a;\text{SO}}_{[\nu,\alpha]} = \hat{P}^b_{\text{en}[\nu,\alpha]} \frac{R_{\alpha\alpha}}{\hat{P}^b_{\alpha\alpha} + R_{\alpha\alpha}}. \tag{30}$$

$$(y^{\mathbf{v}|a;\text{SO}}_\nu - y^{\mathbf{v}|b}_\nu) = \hat{P}^b_{\text{en}[\nu,\alpha]} \frac{y^o_\alpha - y^b_\alpha}{\hat{P}^b_{\alpha\alpha} + R_{\alpha\alpha}} \tag{31}$$

which are the corresponding values that would be obtained if y^o_α was the only assimilated observation. More detailed results are given in Appendix D.

3.2 | Curves in the top graphs

Most central in Figures 2–7 are the blue curves, which are related to the component J^b_α. More precisely, the blue curves in the top graphs of these figures correspond to the sum $S(J^b_\alpha)$ of J^b_α over all observations α which can be found in the respective bins as given on the x-axes of such graphs. From the discussions in Section 2.2.1, it follows that the blue curves have to be greater than zero if the assimilated observations pull the model towards the verifying data. This is a necessary condition for that these observations have a positive impact at all (with respect to this verification metric).

The corresponding reference values $S(\langle J^b_\alpha \rangle_{\text{estim}})$—see Equation (29)—are given by the respective green curves, and in this study the cross-validation mainly refers to the comparison of the blue with the green curves. More precisely, it is a comparison between background error covariances in observation space between the assimilated (y^b_α) and the verifying ($y^{\mathbf{v}|b}_\nu$) observation. The green curves represent the covariance that the DA system employs (which is obtained from the ensemble) and the blue curves show the corresponding values obtained from the observations. Background error covariances are a central part of modern DA systems, which distribute the information from the observations in model space. The cross-validation procedure can be interpreted as a test of how such covariances actually fit with the observations.

The top graphs of Figure 2 also show a curve that is proportional to the traditional EFSOI impact measure J_α. The thin black curves (which correspond to boundaries of the lightly shaded regions in these graphs) yield $2S(J^b_\alpha) - S(J^{ab}_\alpha) = -2S(J_\alpha)$. Important in these graphs are also the curves with the cyan squares, which, as explained in Appendix C, give an impression of the statistical significance of the data collected in the respective bins. These curves give the standard deviation of a corresponding stochastic model process whose increments have the same magnitude as the contributions to the sum $S(J^b_\alpha)$ but for which the sign is completely random (so that the expectation value of this stochastic process is zero). It is clear that if the magnitude of the blue curve is anywhere close to this curve there is a strong possibility that the data in such a bin are not statistically significant. Whereas for strictly Gaussian data the values rarely differ by more than three standard deviations from the truth, we would like to emphasize that the data for which these sums are computed are generally not Gaussian (even in the case that the respective observation and forecast errors are Gaussian), so that encountering more than three standard deviations may not be sufficient to "prove" statistical significance.

3.3 | The normalized (bottom) graphs

The bottom graphs differ from the top graph mainly by the fact that all curves in the bottom graphs are normalized; that is, all curves in these graphs have been divided by the same function N (which is computed independently for each of the respective bins). To facilitate the comparison between the different curves, N has been chosen so that the magnitude of the scaled green curves is always between zero and one. Although in a first trial N was set directly to the values of the green curves, this led to problems in bins where the localization functions $\eta^t(\nu, \alpha)$ are very small and for which the ratios between, for example, the blue and the green curves may assume very large values. Therefore, instead of using Equation (29) we write

$$N = S(\langle \widetilde{J^b_\alpha} \rangle_{\mathrm{estim}}), \tag{32}$$

where

$$\langle \widetilde{J^b_\alpha} \rangle_{\mathrm{estim}} = \sum_\nu \widehat{P}^a_{\mathrm{en}[\nu,\alpha]} \frac{\widetilde{P^b_{\mathrm{en}[\nu,\alpha]}}}{R_{\nu\nu} R_{\alpha\alpha}}$$

is obtained from Equation (29) by replacing $\widehat{P}^b_{\mathrm{en}[\nu,\alpha]}$ through the unlocalized ensemble covariance $\widetilde{P^b_{\mathrm{en}[\nu,\alpha]}}$. Since $\widetilde{P^b_{\mathrm{en}[\nu,\alpha]}}$ differs from $\widetilde{P}^b{}_{\mathrm{en}[\nu,\alpha]}$ only by a factor $\eta^t(\nu, \alpha)$, the green curves in the normalized graphs are effectively weighted averages of the localization function $\eta^t(\nu, \alpha)$—see Equation (B3) for more details—and, therefore, always assume positive values between zero and one.

In addition to most curves from the top graphs, the normalized bottom graphs also show a curve related to the second component J^{ab}_α of J_α. The red curves in the bottom graphs show $S(J^{ab}_\alpha)/N$ and, from the optimality condition, Equation (17), it follows that if the size of the analysis increments was optimal with respect to the verification function J then the red curves should coincide with the respective blue curves.

4 | SOME RESULTS

Most of the statistics presented in the following were produced within the global LETKF, which works in conjunction with the icosahedral nonhydrostatic (ICON) NWP model. Some comparison is also made with the analysis increments from our hybrid ensemble–variational (EnVar) system, which in conjunction with the ICON model provides the best global forecast at DWD. The EnVar system corresponds to a three-dimensional variational (3D-Var) scheme for which the background error covariance matrix is a linear combination of a static component (30%)—which is derived with the National Meteorological Center's method; see Parrish and Derber (1992)—plus a flow-dependent component from the LETKF (70%).[2] The statistics shown in Sections 4.1–4.3 were produced from a test suite at DWD that ran for several months, starting at the beginning of December 2018. More precisely, the data employed correspond to the control of the suite (which is close to DWD's operational system in spring 2019) and data were taken from the 25-day period December 7, 2018, to January 1, 2019. The example in Section 4.4 corresponds to the control run of another test suite that started in March 2020, and the data for this study were taken for the whole month of June 2020.

All the plots presented in this article are for forecast lead time $t = 0$; that is, they indicate the impact

[2] This flow-dependent component proved to be essential for the forecast quality and upgrading the former 3D-Var system to the EnVar system led to a major forecast quality boost at DWD.

on the analysis. This is the simplest case, as it isolates issues regarding the processing of data by the DA system from those related to the initialization and performance of the forecast model (which includes problems related to balances and different types of nonlinearities). Applying the diagnostics to the case $t = 0$ is possible as we use observations for verification—verifying against the analysis requires sufficiently large values of t, as, for example, discussed in Privé *et al.* (2020).

A major objective of this work is to gain some experience of what type of correspondence one may expect for the diagnostics (and their reference values) introduced in the last section. For this we produce such statistics for some-well established observations which are measurements from radiosondes (TEMPs) and aircraft, bending angles from radio occultations (GPSRO) and horizontal wind from AMVs (or satellite observations [SATOBs]). Whereas TEMPs and aircraft provide quasi-independent measurements of temperature, moisture, and horizontal wind, the GPSRO bending angles are sensitive to temperature and (particularly in the troposphere) also to moisture. We always start with the in-situ measurements from TEMPs and aircraft (for which the localization procedure of the LETKF is most appropriate) and use the results as a benchmark for the more complex GPSRO and SATOB measurements.

The new diagnostics introduced in the last section (see Equations 28a,28b, and 29) are all inversely proportional to the observation errors $R_{\alpha\alpha}$ and $R_{\nu\nu}$, which in this context can be regarded as scaling factors. In our global DA system we have to distinguish the observation errors that are provided as input via a namelist (which is determined using the Desroziers method plus some tuning) and the errors that are actually used in the analysis. In our global system (which is true for the hybrid variational and the global LETKF), the latter correspond to the input values plus an increase from the variational quality control (Var QC[3]). In Sections 4.2 and 4.3, the values actually employed are used for the errors $R_{\alpha\alpha}$ of the assimilated observations, and the unmodified input value is used for the verification data. A different choice would have led to some quantitative difference, but the general conclusions of this article would be unchanged.

As already explained, our cross-validation procedure (i.e., comparing the blue and the green curves in Figures 2–7) can be regarded as a test for the correspondence between the background error covariance estimated by the ensemble with that from the observations using Equation (24). In principle, the consistency relations Equations 24 and 25 can be tested in a more straightforward sense (which is, however, less feasible for bin related diagnostics). To give a better impression of the ensemble's overall skill for estimating background error covariances, we start by showing results from such a more direct comparison between the two covariance estimates in subsection 4.1. These are made with the global data set only and are meant to complement the bin related diagnostics. After that, plots showing results for the new diagnostics (as described in Section 3) are presented in Section 4.2, where the first diagnostic test (correspondence between the blue and the green curves) is discussed while Section 4.3 focuses on the diagnostics related to the observational information content of the analysis increments (by comparing the red and the blue curves of those plots). Further, the example of a recent application of the cross-validation method, which allowed identifying bias problems of a small subset of aircraft measurements is presented in Section 4.4.

4.1 | Direct comparison of background error covariances

To get a more direct impression of the ensemble's capability of estimating realistic covariances, we compare the values of the ensemble estimated covariance with the observation-based estimate—using Equation (24) or its generalization Equation (25). In this subsection, all quantities are scaled with the observation error standard deviations that are used as input to our global DA system.[4] Further, for the purpose of this comparison, the localization function has been replaced by a step function, which means that the ensemble covariance without localization $\widetilde{P^b_{\mathrm{en}[\nu,\alpha]}}$, directly obtained via Equation (B1), is taken as the ensemble estimate whereas spurious correlations are (largely) suppressed by limiting the statistics to pairs of observations for which the localization function is larger than 0.5. For producing the respective bottom graphs in Figure 1, these data have first been stratified according to the scaled ensemble covariance

[3] Note that though the purpose of the Var QC is to reduce the weight only for those observations for which the first-guess departure is larger than a specified value (because such observations are less trusted), for the adjoint computations to be continuous a differentiable function is used for the relation between the size of the inflation and the first-guess departures. This means that observation errors are inflated also for small model departures, for which, however, the inflation should be very weak and have no significant influence on the analysis.

[4] Using the errors that are modified by the Var QC would link the scaling factors to the magnitude of the observation minus first-guess departures, which would obscure the direct correspondence of the covariance estimates tested in this subsection.

$$\frac{\widetilde{P^b_{\mathrm{en}[v,\alpha]}}}{\sqrt{R_{vv}R_{\alpha\alpha}}} \tag{33}$$

and then averages have been computed over the closest M neighbors. The x-axis of those graphs corresponds to the averages of the scaled ensemble covariances—mean values of Equation (33) in the respective bins—whereas on the y-axis the average of

$$\frac{(y_v^{\mathbf{v}} - y_v^{\mathbf{v}|b})(y_\alpha^o - y_\alpha^b)}{\sqrt{R_{vv}R_{\alpha\alpha}}} \tag{34}$$

is displayed by the red points (with $M = 500$) and blue lines ($M = 5,000$), which ideally should be on the diagonal (i.e., the black line) if the two covariance estimates corresponded perfectly and the averaging of the observations had fully converged.

However, as argued in Section 2.2, in cases where the variance $\langle (y_\alpha^o - y_\alpha^b)^2 \rangle$ differs significantly from $\hat{P}^b_{\alpha\alpha} + R_{\alpha\alpha}$ (which is the value assumed in the DA system), the value for the covariance employed in the DA system should also differ from what would be expected from Equation (24). Then Equation (25), which can be regarded as a generalization of Equation (24)—and which is more closely related to the optimality condition Equation (17)—becomes more relevant for assessing the suitability of the covariances for the DA process. To allow a test of Equation (25), the cyan curves in Figure 1 show the average values of

$$Q_\alpha \frac{\widetilde{P^b_{\mathrm{en}[v,\alpha]}}}{\sqrt{R_{vv}R_{\alpha\alpha}}} \tag{35}$$

for the respective bins, and if Equation (25) was correct (and the averaging was sufficient and adequate) then the red and blue curves would coincide with the respective cyan curves. Note that if $\langle (y_\alpha^o - y_\alpha^b)^2 \rangle = \hat{P}^b_{\alpha\alpha} + R_{\alpha\alpha}$ holds then one has $Q_\alpha = 1$, in which case the cyan curves are identical to the diagonals in these plots.

To get some indication of the relative importance of the different covariance regimes in the bottom graphs, the black curves in the top graphs are proportional to the number of occurrences $\mathrm{nb}_{\mathrm{bin}}$ of observations pairs $(\boldsymbol{y}_v^{\mathbf{v}}, \boldsymbol{y}_\alpha^o)$ in equidistant covariance bins, whereas the red curve shows the product of $\mathrm{nb}_{\mathrm{bin}}$ with the mean x value (i.e., the mean covariance) of those bins. Since the assimilation impact of an observation is proportional to the size of the covariance (it is zero if the covariance is zero), the actual importance of a covariance regime is better characterized by the red curve.

The impact of assimilating radiosondes on verification against aircraft is shown in Figure 1a,c. For most of the bins, the blue curve in Figure 1c is reasonably close to the black diagonal, which indicates a good to very good correspondence between the covariance estimates from the ensemble and those obtained from the observations. Particularly for large (normalized) covariances, the blue curve is a little below the diagonal which indicates that the ensemble somewhat overestimates the actual covariance. Here, from the fact that the cyan curve is also below the diagonal, the optimality condition—for single-observation experiments, Equation (25)—appears to be reasonably fulfilled for such larger background deviations.

The good correspondence seen from Figure 1c is typical for most in-situ measurements (particularly when both measurements correspond to the same model variable). For zonal wind from AMVs versus aircraft, however, the covariance computed from the observational data is much smaller than the one estimated by the ensemble. This may be related to the known problem of assigning the proper height for AMVs (which means that these observations effectively have a forward operator problem). Note that, as seen from the cyan curve, the problem is partly compensated by having Q_α substantially smaller than one for the AMVs, which is related to observation error inflation.

From the bottom rows of Figure 1 one finds that for observations measuring different quantities, like radiosonde humidity versus aircraft temperature (Figure 1e,g) and GPSRO versus radiosonde temperature (Figure 1f,h), the ensemble's skill for estimating the covariance is a little smaller but generally still quite impressive. An exception are regimes where (in Figure 1g) the ensemble diagnoses positive correlation between aircraft temperature and radiosonde humidity and for which the actual (i.e., observation-based) covariances are seen to be substantially smaller. Though the reasons for the ensemble's overprediction of such positive covariances might be complex, from the top graph one finds that the corresponding weather situations are less frequent and the data seem statistically much less important than those for which negative correlation is diagnosed by the ensemble.

4.2 | Comparing ensemble-based covariances with observations for different types of bins

We first discuss plots regarding in-situ measurements from TEMPs and aircraft. In Figure 2 and Figure 3a–h one can see from the (unnormalized) graphs (Figure 2a–d, i–l, q–t, 3a–d) that the blue and green curves seem to agree reasonably well if the distance between the observations is not too large. This is in accordance with the results from in-situ measurements in Figure 1c, which show a good correspondence between the covariance estimated with the observational data and that from the ensemble. The normalized

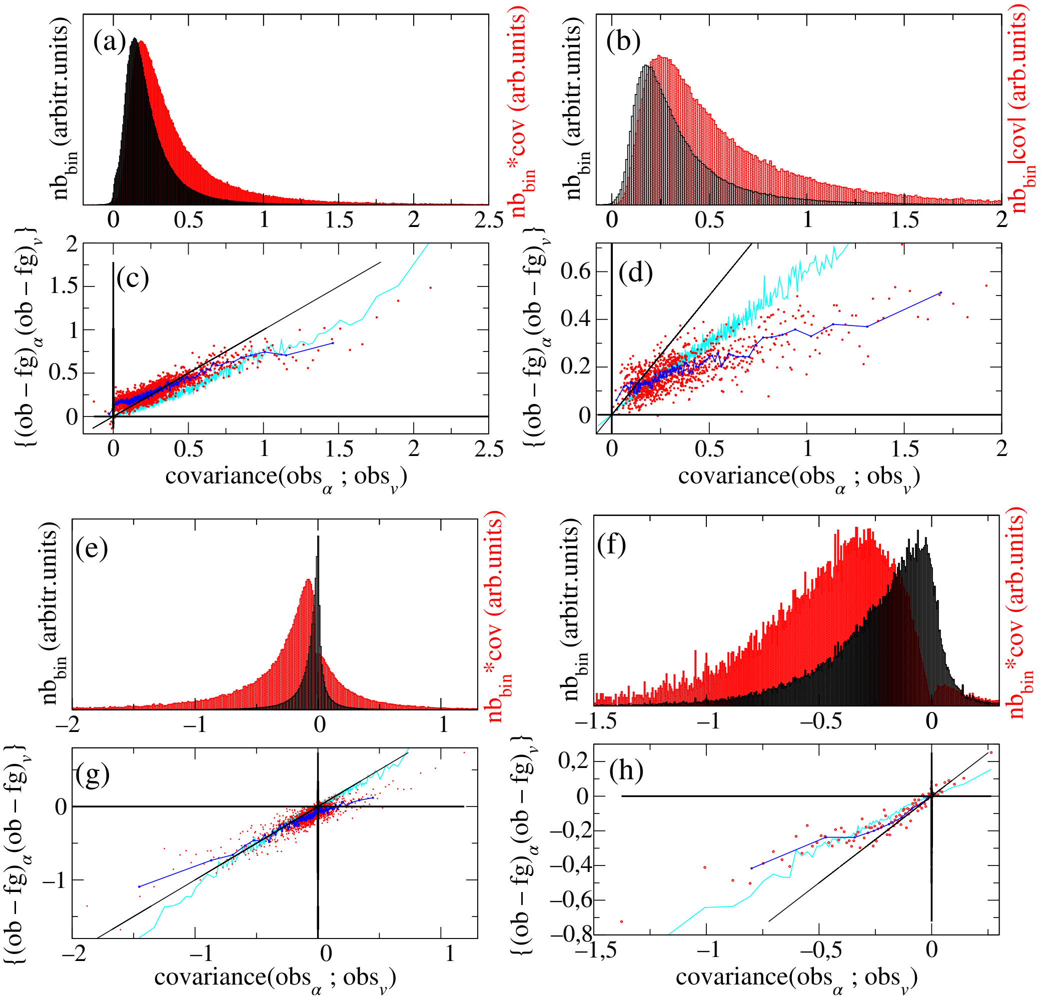


FIGURE 1 Statistics of the background error covariances between different observation types. A comparison is shown between covariance estimates from the ensemble given on the x-axis—value computed from Equation (33)—with that from the observations on the y-axis—as computed from Equation (34. Red dots ($M = 500$) and blue curves ($M = 5,000$) are obtained from averaging of bins of different size M. The cyan curves correspond to the averages over Equation (35). The respective upper graphs (a, b, e, f) show statistics for equally spaced covariance bins as given on the x-axis (the x-axis shows the same quantity as in the bottom graph). Black histogram: number of data nb$_{\text{bin}}$ in bins. Red curve: nb$_{\text{bin}}$abs(cov), where cov is the mean x-value of the respective bins. Units of the curves and the size of the bins in the top graph are arbitrary, as the only purpose is to indicate the relative importance of the respective covariance regimes for the statistics displayed in the respective bottom graph. The top rows show results for zonal wind u with (a, c) u from radiosondes (α) versus u from aircraft (v) and (b, d) u from atmospheric motion vectors (α) versus u from aircraft (v). The bottom rows show (e, g) aircraft temperature (α) versus radiosonde relative humidity (v) and (f, h) GPS radio occultation (α) versus radiosonde temperature (v). ob/obs, observation; fg, first guess [Colour figure can be viewed at wileyonlinelibrary.com]

graphs (Figure 2e–h, m–p, u–x, 3e–h) give a more detailed view and show that, indeed, apart from very few exceptions, the sign of the blue curves is almost everywhere positive, indicating a beneficial impact contribution. In the figures shown in this section, some of the most noisy data bins (with a small data count) have already been discarded by omitting bins at the edge of a plot for which the normalization factor—see Equation (32)—is smaller than one. One finds that for the majority of the plotted bins the blue curve is at least one order of magnitude larger than the magnitude of the corresponding noise estimates (cyan squares), which gives some confidence that the sign of the blue curves in these bins is statistically relevant.

From the normalized bottom graphs, particularly in Figure 2 and the top rows of Figure 3, one can see a general trend that the blue curves tend to increase from north to south (left columns) and from high to low altitudes (i.e., low- to high-pressure levels, second columns, graphs f, n, and v). This probably reflects the larger uncertainty of the model state in the Southern Hemisphere (where we have less observations so that our model state is less constrained) and near the ground (where the atmospheric

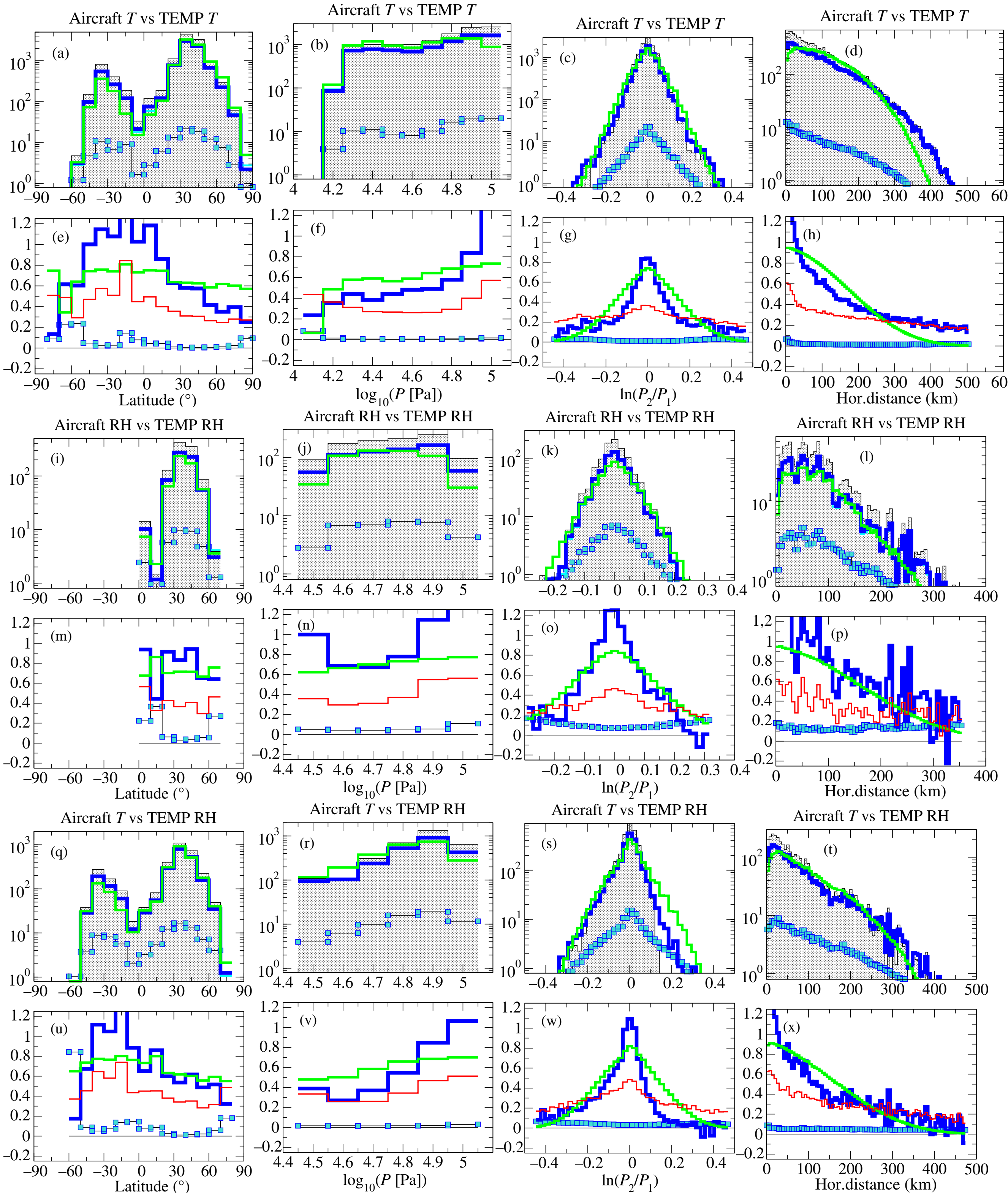


FIGURE 2 Diagnostics described in Section 3 for measurements from aircraft verified by those from radiosondes (TEMP) in bins of (a, e, i, m, q, u) latitude, (b, f ,j, n, r, v) decadic logarithm of vertical pressure (P) coordinate, (c, g, k, o, s, w) difference of (natural) logarithmic pressure levels (positive values: radiosonde is at higher pressure) and (d, h, l, p, t, x) horizontal distance. Statistics are displayed for measurements of (a–h) temperature (T), (i–p) relative humidity (RH), and (r–x) for temperature measurements from aircraft versus humidity measurements from radiosondes [Colour figure can be viewed at wileyonlinelibrary.com]

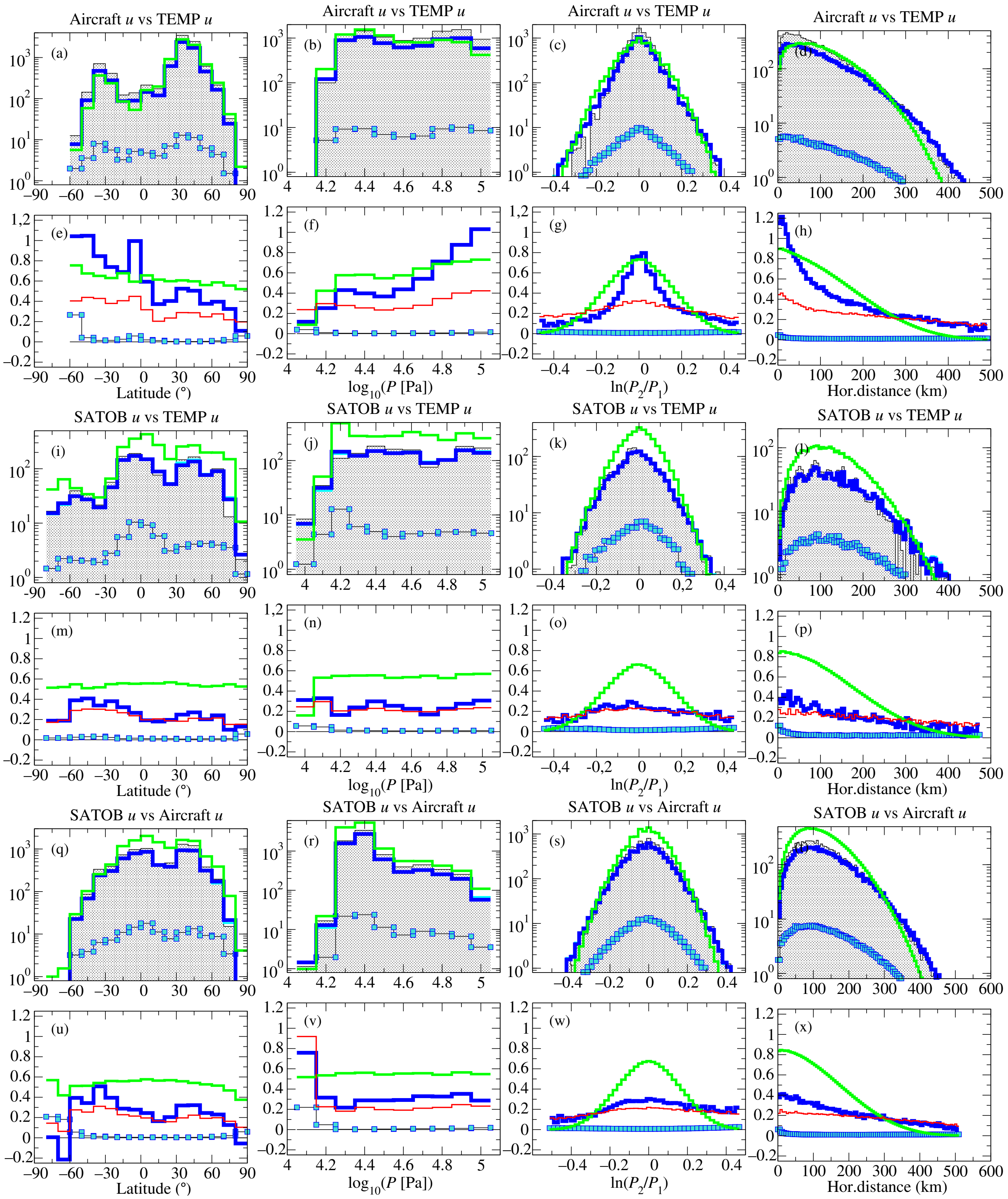


FIGURE 3 The same as Figure 2 but for different measurements of zonal velocity u (instead of temperature and/or humidity). (a–h) u from aircraft verified by radiosondes. (i–x) u from satellite observations (SATOB) versus (i–p) radiosondes (TEMP) and (r–x) aircraft [Colour figure can be viewed at wileyonlinelibrary.com]

flow is more strongly affected by unresolved processes). Larger uncertainty of the model state means that the assimilated observation can have a stronger impact, which explains higher values of the blue curves in these regions. Further, in the two right columns of these figures, the normalized blue curves, like the respective green curves, have distinct peaks at small separation distances. The peaks of the blue curves, however, tend to be higher, indicating that, in these bins, the actual covariance (or variance) of the respective background errors is larger than the values estimated with the ensemble. The peaks of the blue curves, however, decrease much faster in the near range, so that at intermediate distances the blue curves are generally well below the green curves. This underlines the importance of the localization function for reducing the covariances estimated from the ensemble (these results actually suggest that a more rapidly decreasing localization function would lead to even more realistic covariances at intermediate separation distances). At larger separation distances, where the green curves go to zero (proportionally to the localization functions), the blue curves have a much broader tail, which is generally well above the respective green curve.

Results related to the impact of assimilating temperature (from aircraft) on verification against relative humidity (from TEMPs) are shown in the two bottom rows of Figure 2. Though these graphs generally look qualitatively quite similar to those from the upper rows of Figure 2 (where the two independent observations considered in each graph measure the same physical quantity), the most striking difference is probably the asymmetric decay of the blue curve's peak at low vertical distances in Figure 2w. The decay of the blue curve with vertical separation distance is faster when the humidity measurements are at lower altitude (larger pressure) than the temperature measurements. This illustrates that the actual spatial decay of covariances or correlations is more complex than assumed by the localization functions employed. Nevertheless, as already observed in Figure 1g, the ensemble also seems to have significant skill for estimating the cross-covariances between temperature and humidity (and also for other cross-covariances, not shown).

For the zonal wind data from SATOBs displayed in Figure 3i–x, statistics from the two verification types (against TEMP and aircraft) shown in the normalized bottom graphs seem to agree quite well with each other. For both verification types the blue curves are, however, substantially lower than their counterparts from the aircraft versus TEMP statistics in Figure 3a–h. From the normalized graphs in Figure 3o,p,w,x one finds that the blue curves are remarkably flat with respect to the distance between the observations with the small distance peaks typical for in-situ measurements largely missing (or at least strongly reduced) in these data. In contrast to this, the tails at larger vertical distances decay at a similar or even slower pace than for the aircraft versus TEMP diagnostics in Figure 3a–g. Suboptimalities regarding the height assignment for SATOB AMVs have been reported (Folger and Weissmann, 2014; 2016) and the lack of the short distant peak (while the large distance tail is not affected) seems consistent with this. Moreover, looking at the SATOB versus aircraft data (Figure 3w), the tail of the blue curve at positive logarithmic distance, $\ln(P_2/P_1)$, is particularly broad and exhibits higher values than for the aircraft vs TEMP data in Figure 3g. These bins with positive $\ln(P_2/P_1)$ correspond to data where the height of the aircraft measurements is lower than the height assigned to the AMV observations, and having larger than expected covariances in these bins could be related to AMVs for which the assigned altitude is too high, which appears consistent with findings from Folger and Weissmann (2014).

The corresponding statistics for GPSRO bending angles validated against temperature measurements from aircraft are shown in Figure 4a–h. Similar to the results from in-situ data, in the normalized GPSRO related graphs the blue curves show smaller values in the Northern Hemisphere than in the Southern Hemisphere. In stark contrast to in-situ measurement results (as seen in Figure 2 and Figure 3a–h), however, is the dependence on the vertical separation distance where the distinct peak of the blue curves (found for in-situ measurements) seems to be missing in Figure 4g. Instead, there is a substantially flatter and broader maximum peaking at small positive values and some noisy structure at negative separation distances (positive vertical distances indicate that the nominal height assigned to the GPSRO measurement is at greater altitude than the radiosonde). This underlines the more complex structure of the GPSRO measurements, which are non-local, so that the measurement height or localization height (which is assigned by the LETKF to all observations) has to be interpreted as some kind of vertical average of the region that contributes to the measurement. Still, for this non-local observation type also, FSOI statistics show clearly positive benefits with a magnitude that is only a little smaller than our theoretical expectations.

Figure 4i–p shows the corresponding results for verification with temperature measurements from radiosondes. Compared with aircraft data, verification against TEMP generally yields somewhat higher values for the blue curves in most normalized bins (Figure 4m–p). Differences are most significant in the Tropics (Figure 4e,m) and also for large horizontal separation lengths (Figure 4h,p)). From the pressure-level-related bins (Figure 4n) one finds a quite excellent agreement between blue and green curves at higher (i.e., stratospheric) altitudes where one has no aircraft data (as commercial aircraft do not fly there). Indeed, we found that most of the differences between the

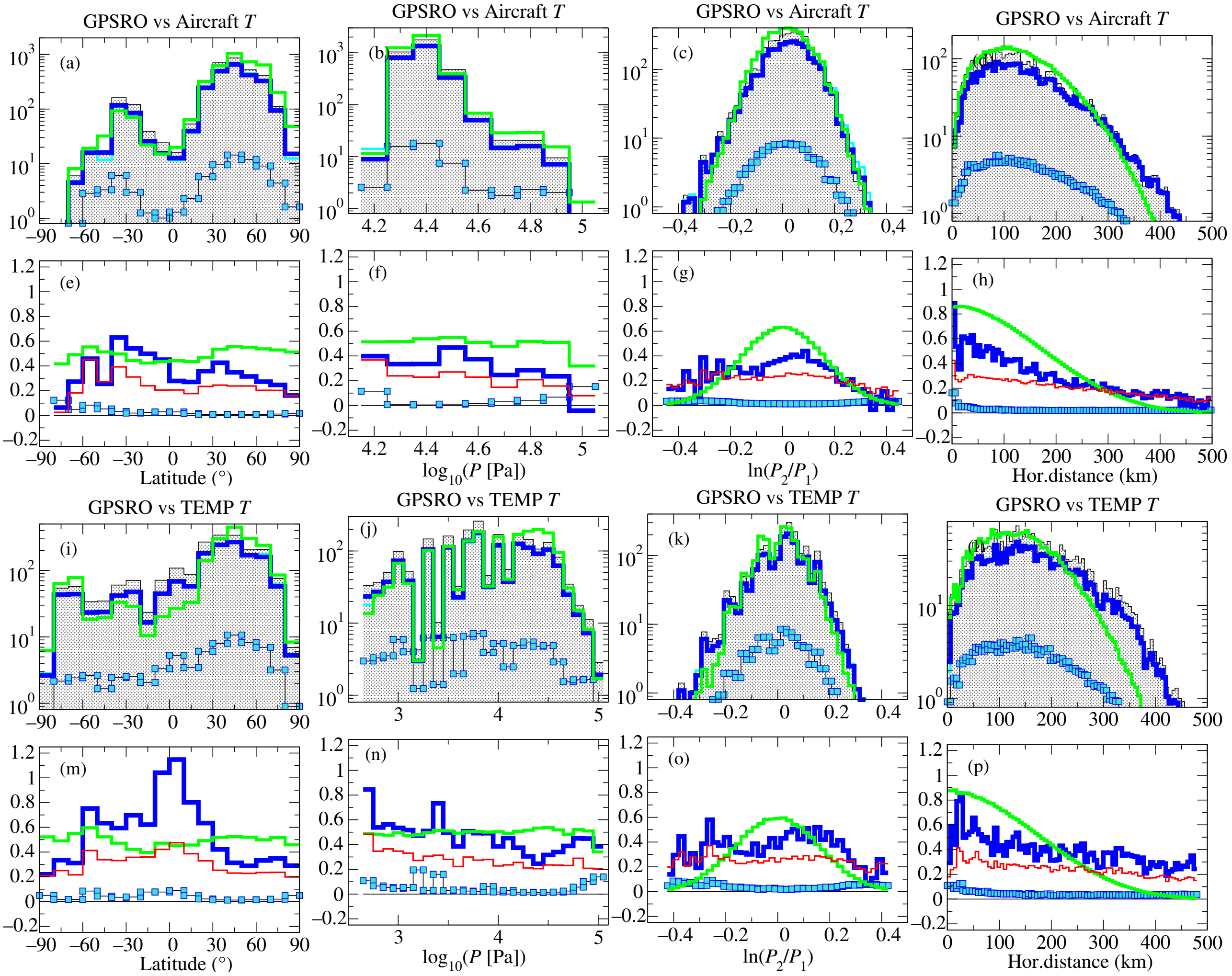


FIGURE 4 The same as Figure 2 but for GPS radio occultation (GPSRO) observations verified by temperature from (a–h) aircraft and (i–p) radiosondes (TEMP) [Colour figure can be viewed at wileyonlinelibrary.com]

top and bottom rows of Figure 4 can be attributed to such altitude; and if restricted to altitudes below 170 hPa (as shown in Figure 5), the differences with what is obtained for aircraft verification diminishes strongly. In particular, the greater horizontal correlation length observed from Figure 4p is mostly related to the stratosphere and much reduced for data from altitudes below 170 hPa (see Figure 5d). This emphasizes the importance of the spatial distribution of the verifying data in those statistics.

4.3 | Information content of analysis increments

In almost all plots discussed so far,[5] red curves are significantly below blue curves for basically all latitude- and height-related bins. The same is true for co-localization-distance-related bins, provided that the distance between the observations is not too large. For the in-situ measurements in particular one finds that the discrepancy between the curves is largest at smaller distances, which might indicate that the DA system is not able to capture the full information content at small scales.

Though this may be partly related to observation-error inflation, we found a strong indication that the finite ensemble size (and the fact that the LETKF constructs the analysis increments only within the quite low-dimensional ensemble space) plays an important role there. To check this point we have reproduced the results from Figure 2 but with the first-guess departures and analysis increments from the LETKF being replaced by those from the hybrid EnVar system[6]. Note that the EnVar

[5] Apart from those related to SATOB data for which the blue curves have particularly small values.

[6] In those graphs, the interpretation of the correspondence between the red and blue curves in terms of the consistency relations still holds even

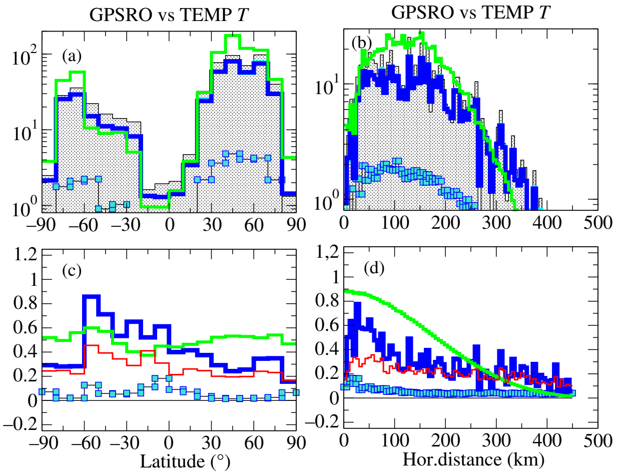


FIGURE 5 The same as the corresponding plots in Figure 4 (i.e., statistics of GPS radio occultation [GPSRO] data verified by radiosonde temperature [TEMP T] in latitude and horizontal distance bins) but for radiosonde data restricted to altitudes lower than 170 hPa [Colour figure can be viewed at wileyonlinelibrary.com]

system uses exactly the same observation errors as the LETKF. The main difference is that the analysis of the EnVar system is not limited to the relatively small ensemble space as for the LETKF.

Comparing Figure 7 with Figure 2a–h, the better correspondence between the red and blue curves for the EnVar increments (shown in Figure 7) is quite striking (particularly at smaller distances between the observations), which seems to confirm the hypothesis that the limitations of the LETKF are a major reason for the comparably low information content of the analysis increments. Such limitations are discussed by Hotta and Ota (2021), who show that the ensemble size gives an upper limit to the information content that can be represented by the analysis increments. These arguments can explain why the larger information content at small distances is not seen in the LETKF analysis increments, whereas the lower information content at large distances appears to be captured similarly well as in the EnVar system (which does not have these limitations).

Note that Figure 7 cannot be used to deduce the actual FSOI score for the EnVar system as the curves were produced using the analysis covariance matrix of the LETKF (and not that of the EnVar system that would be necessary for computing the impact score). It is, however, legitimate to link the different curves to the respective consistency relations discussed earlier, which allows the interpretation that the better fit of the red with the blue curve (compared with Figure 2) indicates a more optimal exploitation of the information content from the assimilated data by the EnVar. This is consistent with the better forecast obtained from the hybrid system.

4.4 | A recent application of the cross-validation method

This subsection gives the example of a recent application of the cross-validation diagnostic. The statistics used for this are from the single-observation version explained in Section 3.1 (and in Appendix D). Apart from the cyan curves in Figures 8 and 10 (which will be explained in the following), the statistics presented in this subsection correspond to their counterparts from original cross-validation diagnostics discussed in Section 4.2.

When cross-validating satellite radiances with radiosondes and aircraft data, for some upper tropospheric channels we found a huge discrepancy between results corresponding to the different verification data. As shown in Figure 8, cross-validation against radiosondes (TEMP) showed quite excellent agreement for Advanced Microwave Sounding Unit (AMSU) channel 7, whereas cross-validation against aircraft data indicated a clearly suboptimal correspondence between the satellite radiance and in-situ temperature measurement. As shown in Figure 9, it was found that the negative contributions to the blue curve in the aircraft verification were dominated by a region over the northwest Atlantic. When restricting statistics to the region marked by the red rectangles in Figure 9 (as is done in Figure 10a), the correspondence between temperature measurements from aircraft and radiosondes turned out to be much worse than the global statistics shown in Figure 2b,f. It turned out that this bad correspondence was caused only by one specific type of aircraft data, often referred to as Aircraft Reports (AIREP)data (which is labeled with code type 141 in the bufr data files), whereas the correspondence of the Aircraft Meteorological Data Relay (AMDAR) temperature measurements in this region was actually quite good (see Figure 10b, where AIREP data have been excluded). As seen from Figure 10c, the AIREP temperature data generally verify very badly against radiosondes. Figure 11 shows that most of the contributions (and particularly the negative/detrimental contributions) from this data type originate from the North Atlantic (east and west), which is therefore likely to have a significant impact on the forecast for Europe.

though the curves in these graphs are not directly linked to the impact of the observations in the EnVar system. Note that, in this figure, the green curves are not true estimators for the blue curves as the background error estimate from the ensemble was used, which accounts only for part of the corresponding covariance from the EnVar system.

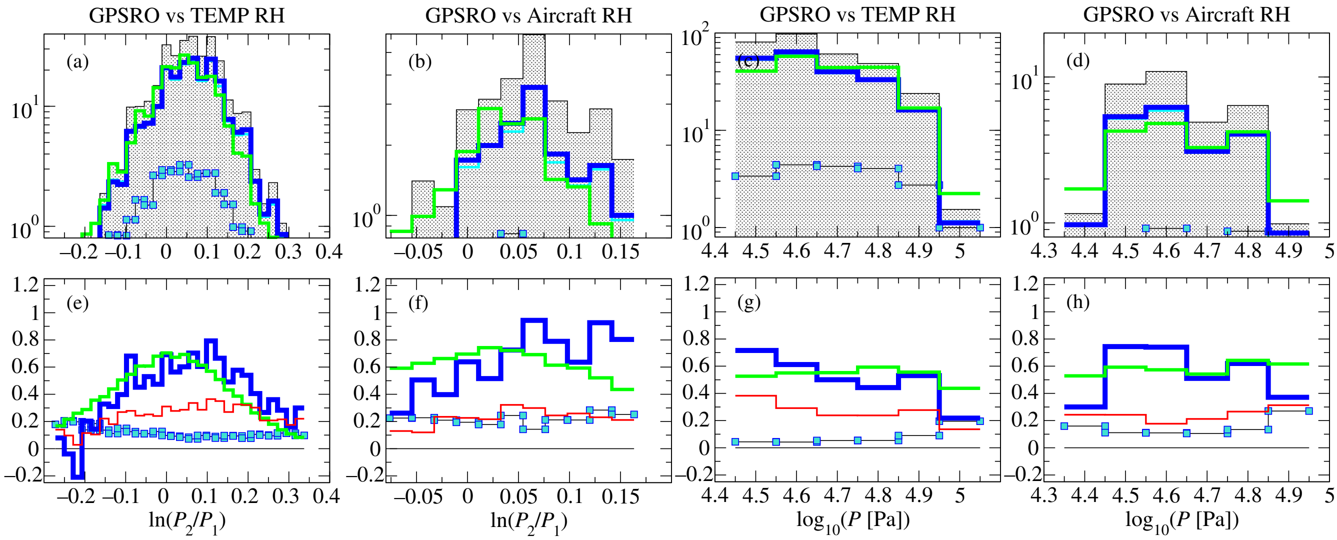


FIGURE 6 The same as the corresponding graphs from Figure 4 but for GPS radio occultation (GPSRO) data verified by relative humidity measurements from (a, c, e, g) radiosondes (TEMP RH) and (b, d, f, h) aircraft measurements. Statistics are collected in bins of (a, b, e, f) logarithmic pressure differences and (c, d, g, h) the logarithmic pressure height of the radiosonde measurement [Colour figure can be viewed at wileyonlinelibrary.com]

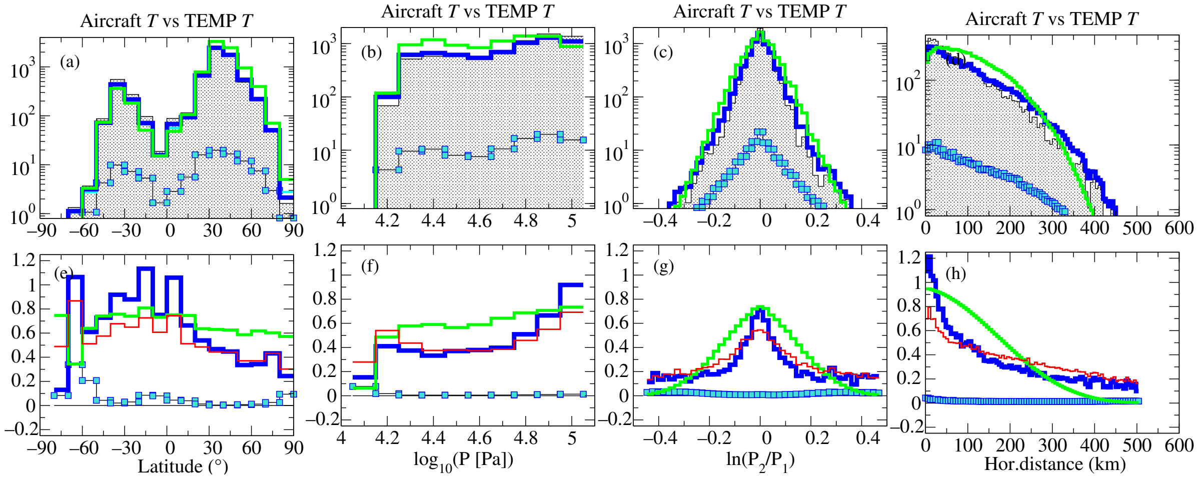


FIGURE 7 The same as the top rows from Figure 2 but replacing the first-guess departures and analysis increments from the localized transform ensemble Kalman filter by the corresponding quantities from the deterministic hybrid ensemble–variational system [Colour figure can be viewed at wileyonlinelibrary.com]

Figure 10a,c actually indicates that the problem with this data type is bias related. The cyan curves in these graphs are taken from the same statistics as the blue curves. The only difference is that for each of the pressure bins the mean bias of the bin has been subtracted from the aircraft measurements. A bias problem of the AIREP temperature data is plausible, as in contrast to the corresponding AMDAR data the AIREP temperatures are not bias corrected since the aircraft temperature bias correction (as implemented in our DA system) requires some aircraft identification, which is not provided for this data type.

Following the discovery of this bias problem, data denial experiments were performed. These showed, mostly for Europe, a small but consistent forecast improvement when the AIREP temperature data where not assimilated. As a consequence, these data have been black listed in the DWD's operational system. It should, however, be noted that the clearly positive (beneficial) values of the cyan curves in Figure 10a,c suggest that a positive impact from these data should be achievable by some relatively simple bias correction method. We therefore expect to be able to put these data back into operations after adequate changes to the bias correction procedure for aircraft

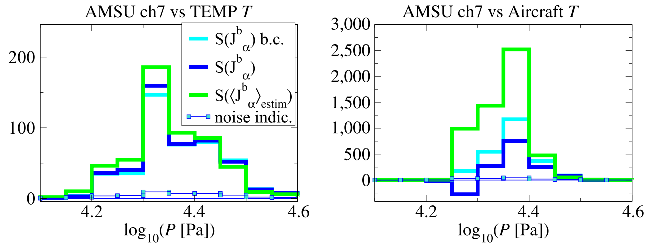


FIGURE 8 Cross-validation between brightness temperature from the Advanced Microwave Sounding Unit (AMSU) A channel 7 with temperature measurements from (a) radiosondes and (b) aircraft. The curves correspond to the same kind of statistics as, for example, in Figure 2b but for the single-observation version of the cross-validation diagnostics (see Sections 3.1 and 3.2). Shown is also the bias corrected curve "$S(J^b_\alpha)$ b.c." which corresponds to the sum $S(J^b_\alpha)$ over the terms J^b_α but for which, in each pressure bin, the mean (i.e., the bias) of the brightness temperatures' first-guess departures has been subtracted from the satellite data. [Colour figure can be viewed at wileyonlinelibrary.com]

temperature data (efforts in this direction are currently under way).

5 | SUMMARY AND FURTHER DISCUSSION

This work shows that the standard FSOI diagnostic that is aimed at quantifying assimilation-related forecast improvements can be written as the sum of two diagnostic components, each giving insight into a different part of the DA process. More precisely, the question of whether the assimilation of a given observation type improves the fit to some verification data is partitioned into the following sub-questions:

1. Do the first-guess departures (when processed with the ensemble-based covariances) pull the analysis state in the direction of the verifying data?
2. Do the analysis increments contain the observational information in an optimal way (with respect to the verification function J)?

It is clear that the first point (addressed by the first diagnostic component) is most fundamental for assessing whether an observation type could have a beneficial impact at all (given the employed background error covariances). In contrast, the size of the analysis increments that is assessed by the second diagnostic (related to the second point) is sensitive to the relative weight given to the background and observations, and more generally to the functioning and fine-tuning of the DA solver. To this end, we found that, in our global NWP system, the analysis increments produced by the hybrid EnVar system are significantly more optimal than those from the LETKF, which we believe to be due to the small dimension of the ensemble space to which the LETKF is limited when searching for the cost function minimum.

Most of this article focused on the first question, and it is a major objective of this work to establish a cross-validation formalism based on the consistency relation corresponding to the first diagnostic component. The aim is the testing of new or less-trusted observation types by cross-validation with more established ones. The method presented is much more flexible than the testing through denial experiments, allowing the cross-validation of individual observation types (e.g., like AMVs from certain latitudes or vertical heights, or with other specified conditions), which may be essential for finding the source of a suboptimality or for understanding the outcome of testing or tuning measures that are designed to improve

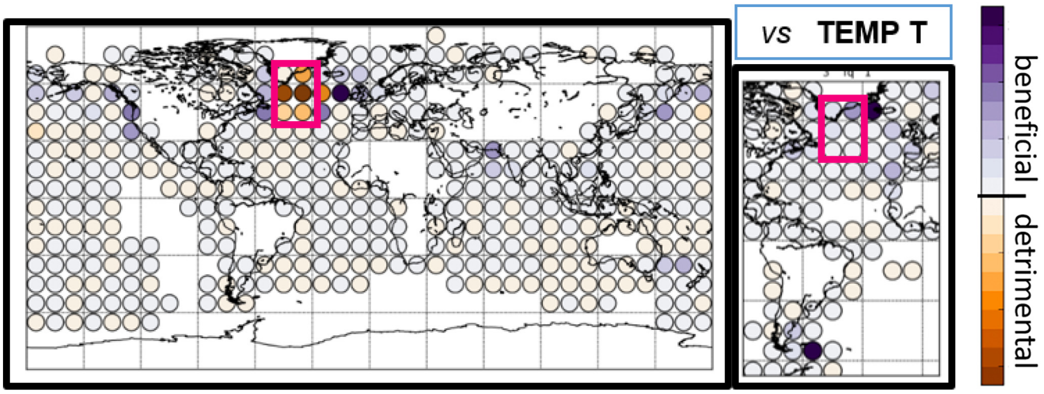


FIGURE 9 Geographical distributions of the contributions to the su $S(S^b_\alpha)$ over the terms J^b_α in Figure 8 for Advanced Microwave Sounding Unit A channel 7 brightness temperature verified by (left) aircraft and (right) radiosondes. Positive (negative) contributions are beneficial (detrimental) and indicate that the assimilation of the satellite data pulls the model closer towards (further away from) the corresponding verification data. [Colour figure can be viewed at wileyonlinelibrary.com]

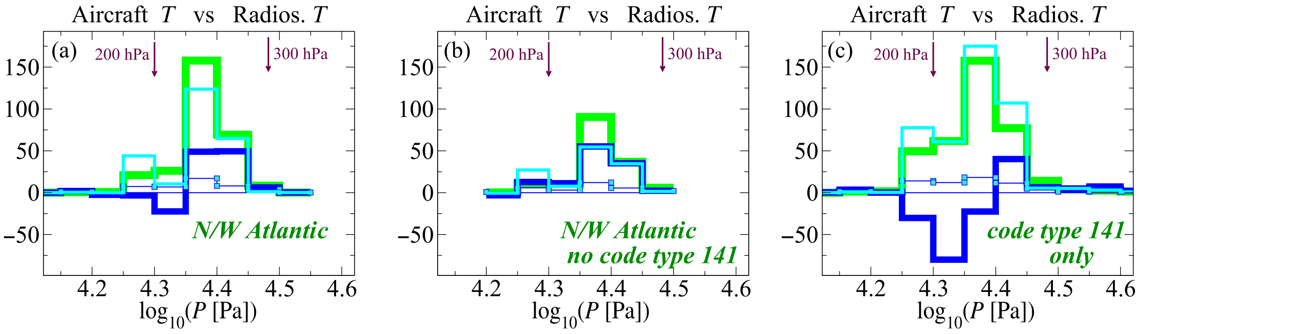
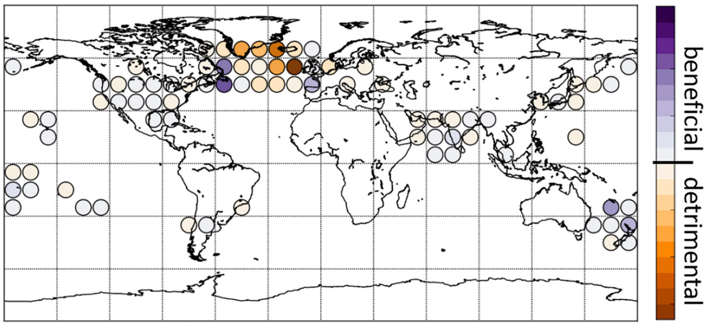


FIGURE 10 The same cross-validation diagnostics as in Figure 8 but for temperature measurements from aircraft validated against radiosondes for (a) a region over the northwest Atlantic (40°–70° N × 30°–50° E), (b) like (a) but Aircraft Reports (AIREP) data (code type 141) excluded, and (c) no spatial restrictions but for AIREP data only. Statistics have been collected in logarithmic bins related to the vertical pressure levels of the radiosonde measurements as indicated on the x-axis. [Colour figure can be viewed at wileyonlinelibrary.com]

FIGURE 11 The same geographical distribution as plotted in Figure 9 but for temperature data from Aircraft Reports (code type 141) verified by radiosonde temperature as considered in Figure 10. [Colour figure can be viewed at wileyonlinelibrary.com]



the impact of some observation type. For this, the identification of biased aircraft data described in Section 4.4 is a good example.

The proposed method depends strongly on the suitability of the (ensemble-based) background error covariances employed, and an important part of this work was to produce such cross-validation diagnostics with observations that are of good quality (and which are known to have a clearly positive impact). For this, the independent in-situ measurements from TEMPs and aircraft played a central role. Also, a more direct assessment of the ensemble estimated covariances has been presented in this article, which, however, was performed on the whole global dataset (and not related to individual bins).

Overall, tests with in-situ observations only, showed an excellent agreement of the proposed consistency relations, though some limitations of the ensemble-based covariances could be identified. For example, the decay of ensemble covariances with collocation distance was found to be slower than for the observation-based covariances. Also, the decay with the vertical collocation distance of the observation-based cross-covariances of humidity and temperature showed some asymmetry that is not reproduced by the ensemble-based covariance estimates. Still, the overall excellent agreement with the consistency relation for the in-situ measurements gives us some confidence that the ensemble-based covariances are of sufficient quality to allow a meaningful interpretation of the proposed cross-validation diagnostics.

Apart from giving some impression of the quality of the ensemble-estimated covariances, these results from in-situ observations also served as a benchmark for testing the more complex observation of GPSRO bending angles and SATOB winds. This particularly allowed us to identify some less-optimal (though still clearly beneficial) behavior of the SATOB wind data that appears consistent with problems related to the height assignment reported by Folger and Weissmann (2014; 2016).

For GPSRO, the agreement with in-situ measurements is overall quite good, particularly when verified with radiosonde temperature in the stratosphere, where the horizontal correlation length for background errors is much larger than at lower vertical levels. A marked differences of GPSRO to the case of in-situ measurements (both verified against other in-situ measurements) is that the covariance has a more complex dependence on the vertical distance between the observations. This underlines the non-local nature of the bending angle measurements. The results suggest that somewhat lower vertical levels (i.e., localization heights) than the ones currently assigned for the GPSRO data might actually lead to a slightly higher FSOI score particularly when verified against humidity measurements.

The data discussed in this article (and particularly the example in Section 4.4) illustrate that, though requiring a lot of care (as for any statistical investigation), the interpretation of the proposed diagnostics in terms of consistency relations can give some valuable insight into how observational data are processed in the DA system and can help to identify situations or aspects that are less optimal than one might assume.

5.1 | Future work and applications

The new diagnostics have a great wealth of potential applications for which the work presented is meant as a starting point that will help the interpretation of these diagnostics when applied to more complex situations and observation types. Depending on the investigation, this may involve further related diagnostics, some of which, in the framework presented here, can be constructed in a straightforward manner. Examples of this are the bias-subtracted diagnostic given by the cyan curves in Figure 10 that helped tracing back the bad cross-validation results of some aircraft data to a bias problem. Also, multidimensional bins (like the latitude–longitude bins in the geographical plots in Figures 9 and 11) may be important for some investigations.

An obvious extension of the results presented herein is to produce statistics for forecast lead times $t > 0$. This involves covariances at different times (t_0 and $t_0 + t$) and comparing the results with those for $t = 0$ presented herein, which may give some insight into the ability of the forecast model to propagate observational information in time. This may give some indication of which aspects of observational information are most relevant for producing a good forecast (rather than a good analysis, which was the focus of the work presented herein). Such results may be affected by imbalances of the analysis state (e.g., resulting from feedback with model parametrizations) or other model issues that make their interpretation more challenging so that the comparison with the results for $t = 0$ presented herein (which are not affected by these influences) may be of great relevance.

A major observation type in modern DA are satellite radiances. Their great importance, together with their sensitivity to various (possibly detrimental) influences, like undetected clouds, aerosols, or trace gases, makes the application of this method to this data type one of our priorities. Applying the new ensemble-based diagnostics to satellite radiances has two aspects. One is the processing in the LETKF, and particularly the impact of localization on these strongly non-local data (whose assimilation in an ensemble system with vertical localization is always to some degree suboptimal). The other is the identification of issues related directly to the observations and their first-guess departures. In practice, we are most interested in observation-related issues, particularly as they should also be relevant to our hybrid EnVar system, which also makes use of the ensemble-estimated covariance matrix and which produces DWD's best global forecast. Detecting such issues is, however, strongly dependent on an adequate localization, as the ensemble-based diagnostics compare observations only within the assigned localization region. Therefore, as will be described in a future publication, finding a good localization height for satellite radiances has been a first step for investigating this observation type with the new diagnostics.

We would like to point out that we expect the probably most widespread application of our cross-validation method to be in the context of single-observation impact assessment (described in Section 3.1 and Appendix D), which was also employed in the example given in Section 4.4. The single-observation diagnostics exploit the same consistency relation as the cross-validation diagnostic related to the full analysis but do not require any input related to the analysis state and, therefore, allow the testing of alternative preprocessing procedures without repeating the analysis step. For example, alternative cloud screening methods, changes to height assignments, or, generally, any changes to the observation operator can be tested with respect to their consistency with the verifying observations. It particularly permits the testing of new observations prior to their use in the assimilation system.

AUTHOR CONTRIBUTIONS

Olaf Stiller: conceptualization; formal analysis; investigation; methodology; project administration; software; validation; visualization; writing – original draft; writing – review and editing.

ACKNOWLEDGMENTS

I would like to thank my colleagues from DWD for interesting discussions and inspirations, particularly Roland Potthast and Christoph Schraff, as well as Christina Köpken-Watts, Harald Anlauf, and Michael Denhard, who gave me some very constructive comments on versions of this manuscript. The final version of this article benefited greatly from the insightful and dedicated critiques of two anonymous reviewers, to whom I am greatly indebted. Open Access funding enabled and organized by Projekt DEAL.

ORCID

Olaf Stiller https://orcid.org/0000-0001-7953-7393

REFERENCES

Buehner, M., Du, P. and Bédard, J. (2018) A new approach for estimating the observation impact in ensemble–variational data assimilation. *Monthly Weather Review*, 146(2), 447–465.

Cardinali, C. (2018) Forecast sensitivity observation impact with an observation-only based objective function. *Quarterly Journal of the Royal Meteorological Society*, 144(716), 2089–2098.

Chen, T.-C. and Kalnay, E. (2019) Proactive quality control: Observing system simulation experiments with the Lorenz '96 model. *Monthly Weather Review*, 147(1), 53–67.

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

Folger, K. and Weissmann, M. (2014) Height correction of atmospheric motion vectors using satellite lidar observations from CALIPSO. *Journal of Applied Meteorology and Climatology*, 53(7), 1809–1819.

Folger, K. and Weissmann, M. (2016) Lidar-based height correction for the assimilation of atmospheric motion vectors. *Journal of Applied Meteorology and Climatology*, 55(10), 2211–2227.

Gaspari, G. and Cohn, S.E. (1999) Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society*, 125(554), 723–757.

Gasperoni, N.A. and Wang, X. (2015) Adaptive localization for the ensemble-based observation impact estimate using regression confidence factors. *Monthly Weather Review*, 143(6), 1981–2000.

Hollingsworth, A. and Lönnberg, P. (1986) The statistical structure of short-range forecast errors as determined from radiosonde data. Part I: The wind field. *Tellus A*, 38(2), 111–136.

Hotta, D., Chen, T.-C., Kalnay, E., Ota, Y. and Miyoshi, T. (2017a) Proactive QC: A fully flow-dependent quality control scheme based on EFSO. *Monthly Weather Review*, 145(8), 3331–3354.

Hotta, D., Kalnay, E., Ota, Y. and Miyoshi, T. (2017b) EFSR: Ensemble forecast sensitivity to observation error covariance. *Monthly Weather Review*, 145(12), 5015–5031.

Hotta, D. and Ota, Y. (2021) Why does ENKF suffer from analysis overconfidence? An insight into exploiting the ever-increasing volume of observations. *Quarterly Journal of the Royal Meteorological Society*, 147(735), 1258–1277.

Houtekamer, P.L. and Zhang, F. (2016) Review of the ensemble Kalman filter for atmospheric data assimilation. *Monthly Weather Review*, 144(12), 4489–4532.

Hunt, B.R., Kostelich, E.J. and Szunyogh, I. (2007) Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Physica D: Nonlinear Phenomena*, 230(1-2), 112–126.

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

Kotsuki, S., Kurosawa, K. and Miyoshi, T. (2019) On the properties of ensemble forecast sensitivity to observations. *Quarterly Journal of the Royal Meteorological Society*, 145(722), 1897–1914.

Langland, R.H. and Baker, N.L. (2004) Estimation of observation impact using the NRL atmospheric variational data assimilation adjoint system. *Tellus A: Dynamic Meteorology and Oceanography*, 56(3), 189–201.

Lien, G.-Y., Hotta, D., Kalnay, E., Miyoshi, T. and Chen, T.-C. (2018) Accelerating assimilation development for new observing systems using EFSO. *Nonlinear Processes in Geophysics*, 25(1), 129–143.

Liu, J. and Kalnay, E. (2008) Estimating observation impact without adjoint model in an ensemble Kalman filter. *Quarterly Journal of the Royal Meteorological Society*, 134(634), 1327–1335.

Ménard, R. and Deshaies-Jacques, M. (2018a) Evaluation of analysis by cross-validation. Part I: Using verification metrics. *Atmosphere*, 9(3), 86

Ménard, R. and Deshaies-Jacques, M. (2018b) Evaluation of analysis by cross-validation, part II: Diagnostic and optimization of analysis error covariance. *Atmosphere*, 9(2), 70

Necker, T., Weissmann, M. and Sommer, M. (2018) The importance of appropriate verification metrics for the assessment of observation impact in a convection-permitting modelling system. *Quarterly Journal of the Royal Meteorological Society*, 144(714), 1667–1680.

Parrish, D. and Derber, J. (1992) The National Meteorological Center's spectral statistical–interpolation analysis system. *Monthly Weather Review*, 120, 1747–1763.

Privé, N., Errico, R.M., Todling, R. and El Akkraoui, A. (2020) Evaluation of adjoint-based observation impacts as a function of forecast length using an observing system simulation experiment. *Quarterly Journal of the Royal Meteorological Society.*

Sommer, M. and Weissmann, M. (2016) Ensemble-based approximation of observation impact using an observation-based verification metric. *Tellus A: Dynamic Meteorology and Oceanography*, 68(1), 27885

Todling, R. (2013) Comparing two approaches for assessing observation impact. *Monthly weather review*, 141(5), 1484–1505.

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

APPENDIX A. THE OPTIMALITY CONDITION

A.1 Condition for a global minimum of the verification function *J*

Here, we show the equality in the upper line of Equation (21) that proves the optimality condition in Equation (17) is a necessary condition for $\boldsymbol{x}^a$ being a global minimum of the verification function J. Using the definition of the metric and the notation

$$\boldsymbol{y}^{\mathrm{v}|\hat{a}} = \boldsymbol{y}^{\mathrm{v}|a} + \delta\boldsymbol{y}^{\mathrm{v}|a}$$

with

$$\delta\boldsymbol{y}^{\mathrm{v}|a} = \delta\lambda_\alpha^{(\mathrm{v})}\boldsymbol{H}^{\mathrm{v}}\boldsymbol{M}_t\mathbf{K}\boldsymbol{\Pi}_\alpha(\boldsymbol{y}^o - \boldsymbol{y}^b)$$

one can write

$$\begin{aligned}
\langle\|\boldsymbol{y}^{\mathrm{v}} - \boldsymbol{y}^{\mathrm{v}|\hat{a}}\|^2\rangle &= \langle[(\boldsymbol{y}^{\mathrm{v}} - \boldsymbol{y}^{\mathrm{v}|a}) - \delta\boldsymbol{y}^{\mathrm{v}|a}]^{\mathrm{T}} \\
&\quad \times \boldsymbol{C}^{-1}[(\boldsymbol{y}^{\mathrm{v}} - \boldsymbol{y}^{\mathrm{v}|a}) - \delta\boldsymbol{y}^{\mathrm{v}|a}]\rangle \\
&= \langle\|\boldsymbol{y}^{\mathrm{v}} - \boldsymbol{y}^{\mathrm{v}|a}\|^2\rangle + \langle\|\delta\boldsymbol{y}^{\mathrm{v}|a}\|^2\rangle \\
&\quad - 2\langle(\boldsymbol{y}^{\mathrm{v}} - \boldsymbol{y}^{\mathrm{v}|a})^{\mathrm{T}}\boldsymbol{C}^{-1}\delta\boldsymbol{y}^{\mathrm{v}|a}\rangle.
\end{aligned}$$

To see that this equals the right-hand side of the first line of Equation (21), we rewrite the expression in the last angle brackets on the right-hand side using

$$\begin{aligned}
&(\boldsymbol{y}^{\mathrm{v}} - \boldsymbol{y}^{\mathrm{v}|a})^{\mathrm{T}}\boldsymbol{C}^{-1}\delta\boldsymbol{y}^{\mathrm{v}|a} \\
&\quad = \delta\lambda_\alpha^{(\mathrm{v})}\{(\boldsymbol{y}^{\mathrm{v}} - \boldsymbol{y}^{\mathrm{v}|a})^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{H}^{\mathrm{v}}\boldsymbol{M}_t\mathbf{K}\boldsymbol{\Pi}_\alpha(\boldsymbol{y}^o - \boldsymbol{y}^b)\} \\
&\quad = \delta\lambda_\alpha^{(\mathrm{v})}\{J_\alpha^b - J_\alpha^{ab}\} \\
&\quad = (\delta\lambda_\alpha^{(\mathrm{v})})^2\|\boldsymbol{H}^{\mathrm{v}}\boldsymbol{M}_t\mathbf{K}\boldsymbol{\Pi}_\alpha(\boldsymbol{y}^o - \boldsymbol{y}^b)\|^2,
\end{aligned}$$

where in the second line the definitions from Equations 4a and 4b have been used, whereas the last line follows directly from the definition of $\delta\lambda_\alpha^{(\mathrm{v})}$ in Equation (19).

A.2 Condition for a local minimum of the verification function J

From the arguments in Section 2.2 it is clear that the optimality condition, Equation (17), has to hold for any global minimum of the verification function J. In the following, we show that this is the case also for any local minimum. More precisely, we show that Equation (17) is a necessary condition for that the derivative of J with respect to the initial conditions is zero (at a local extremum of a differentiable function the derivative is zero in all directions). For this, we replace the analysis $\boldsymbol{x}^a$ by the more general initial conditions $\boldsymbol{x}^{\mathrm{ini}}$ with

$$\boldsymbol{x}^{\mathrm{ini}}(\{\lambda_\alpha\}) = \boldsymbol{x}^{\mathrm{b}} + \sum_{\alpha\in\{\mathrm{obs}\}}\mathbf{K}\lambda_\alpha\boldsymbol{\Pi}_\alpha(\boldsymbol{y}^o - \boldsymbol{y}^b),$$

which coincide with $\boldsymbol{x}^a$ if all the scaling factors λ_α equal one. With this we generalize the verification function by replacing $\boldsymbol{y}^{\mathrm{v}|a}$ by $\boldsymbol{y}^{\mathrm{v}|\mathrm{ini}}$, which is the model equivalent to $\boldsymbol{y}^{\mathrm{v}}$ based on a model run starting from $\boldsymbol{x}^{\mathrm{ini}}$. One has

$$J[\boldsymbol{x}^{\mathrm{ini}}] = \frac{1}{2}(\|\boldsymbol{y}^{\mathrm{v}} - \boldsymbol{y}^{\mathrm{v}|\mathrm{ini}}\|^2 - \|\boldsymbol{e}^b\|^2)$$

and

$$\begin{aligned}
\frac{\boldsymbol{d}(J[\boldsymbol{x}^{\mathrm{ini}}])}{\boldsymbol{d}\lambda_\alpha} &= (\boldsymbol{y}^{\mathrm{v}} - \boldsymbol{y}^{\mathrm{v}|\mathrm{ini}})^{\mathrm{T}}\boldsymbol{C}^{-1}\frac{\boldsymbol{d}(\boldsymbol{y}^{\mathrm{v}|\mathrm{ini}})}{\boldsymbol{d}\lambda_\alpha} \\
&= (\boldsymbol{y}^{\mathrm{v}} - \boldsymbol{y}^{\mathrm{v}|\mathrm{ini}})^{\mathrm{T}}\boldsymbol{C}^{-1}\boldsymbol{H}^{\mathrm{v}}\boldsymbol{M}_t\mathbf{K}\boldsymbol{\Pi}_\alpha(\boldsymbol{y}^o - \boldsymbol{y}^b), \qquad \text{(A1)}
\end{aligned}$$

where we have used

$$\begin{aligned}
\frac{\boldsymbol{d}(\boldsymbol{y}^{\mathrm{v}|\mathrm{ini}})}{\boldsymbol{d}\lambda_\alpha} &= \boldsymbol{H}^{\mathrm{v}}\boldsymbol{M}_t\frac{\boldsymbol{d}(\boldsymbol{x}^{\mathrm{ini}}(\{\lambda_\alpha\}))}{\boldsymbol{d}\lambda_\alpha} \\
&= \boldsymbol{H}^{\mathrm{v}}\boldsymbol{M}_t\mathbf{K}\boldsymbol{\Pi}_\alpha(\boldsymbol{y}^o - \boldsymbol{y}^b).
\end{aligned}$$

Note that for $\boldsymbol{x}^{\mathrm{ini}} = \boldsymbol{x}^a$, Equation (A1) can be written as

$$\left.\frac{\boldsymbol{d}(J[\boldsymbol{x}^{\mathrm{ini}}])}{\boldsymbol{d}\lambda_\alpha}\right|_{\{\lambda_\alpha\}=\{1,1,\dots,1\}} = J_\alpha^b - J_\alpha^{ab},$$

so that Equation (17) follows directly from

$$\left.\frac{\boldsymbol{d}\langle J[\boldsymbol{x}^{\mathrm{ini}}]\rangle}{\boldsymbol{d}\lambda_\alpha}\right|_{\boldsymbol{x}^{\mathrm{ini}}=\boldsymbol{x}^a} = 0,$$

which has to hold if $\langle J[\boldsymbol{x}^{\mathrm{ini}}]\rangle$ has a local minimum at $\boldsymbol{x}^{\mathrm{ini}} = \boldsymbol{x}^a$.

A.3 Optimality condition and error covariances

In the following we show that Equation (22) is a sufficient condition for the optimality condition Equation (17). For this, using the definition and properties of the trace function $\mathrm{Tr}[:]$, we write the expectation values of J_α^b and J_α^{ab} (see Equations 4a and 4b) in the form

$$\langle J_\alpha^b\rangle = \mathrm{Tr}[\boldsymbol{C}^{-1}\widehat{\boldsymbol{P}}^a\{\mathrm{v},o\}\boldsymbol{R}^{-1}\boldsymbol{\Pi}_\alpha\langle(\boldsymbol{y}^o - \boldsymbol{y}^b)(\boldsymbol{y}^{\mathrm{v}} - \boldsymbol{y}^{\mathrm{v}|b})^{\mathrm{T}}\rangle] \qquad \text{(A2a)}$$

$$\langle J_\alpha^{ab}\rangle = \mathrm{Tr}[\boldsymbol{C}^{-1}\widehat{\boldsymbol{P}}^a\{\mathrm{v},o\}\boldsymbol{R}^{-1}\boldsymbol{\Pi}_\alpha\langle(\boldsymbol{y}^o - \boldsymbol{y}^b)(\boldsymbol{y}^{\mathrm{v}|a} - \boldsymbol{y}^{\mathrm{v}|b})^{\mathrm{T}}\rangle], \qquad \text{(A2b)}$$

where J_α^{ab} differs from J_α^b only by $\boldsymbol{y}^{\mathrm{v}}$ being replaced by $\boldsymbol{y}^{\mathrm{v}|a}$ in the second line. One finds that Equation (17) holds if

$$\boldsymbol{\Pi}_\alpha\langle(\boldsymbol{y}^o - \boldsymbol{y}^b)(\boldsymbol{y}^{\mathrm{v}|a} - \boldsymbol{y}^{\mathrm{v}|b})^{\mathrm{T}}\rangle = \boldsymbol{\Pi}_\alpha\langle(\boldsymbol{y}^o - \boldsymbol{y}^b)(\boldsymbol{y}^{\mathrm{v}} - \boldsymbol{y}^{\mathrm{v}|b})^{\mathrm{T}}\rangle, \qquad \text{(A3)}$$

and in the following we will show that this is equivalent to Equation (22).

For this we use the Kalman gain matrix of the following form, which is algebraically equivalent to Equation (9):

$$\mathbf{K} = \boldsymbol{P}^b\mathbf{H}^{\mathrm{T}}[\hat{\mathbf{P}}^b + \boldsymbol{R}]^{-1}, \tag{A4}$$

with $\hat{\mathbf{P}}^b = \mathbf{H}\boldsymbol{P}^b\mathbf{H}^{\mathrm{T}}$ being the background covariance matrix $\mathbf{P}^b$ (at analysis time t_0) transformed into observation space (of the assimilated observations). We further define

$$\hat{\boldsymbol{P}}^b\{\mathbf{v}, o\} \equiv \boldsymbol{H}^{\mathbf{v}}\boldsymbol{M}_t\mathbf{P}^b\mathbf{H}^{\mathrm{T}},$$

which (in strict analogy to $\hat{\boldsymbol{P}}^a\{\mathbf{v}, o\}$ in Equation (5)) would yield $\hat{\boldsymbol{P}}^b\{\mathbf{v}, o\} = \mathrm{cov}(\boldsymbol{y}^{\mathbf{v}|b}, \boldsymbol{y}^b)$ if the respective linear operators and covariance matrices used by the NWP system were fully valid. With this we can write

$$(\boldsymbol{y}^{\mathbf{v}|a} - \boldsymbol{y}^{\mathbf{v}|b}) = \hat{\boldsymbol{P}}^b\{\mathbf{v}, o\}[\hat{\mathbf{P}}^b + \boldsymbol{R}]^{-1}(\boldsymbol{y}^o - \boldsymbol{y}^b), \tag{A5}$$

and therefore

$$\begin{aligned}&\langle(\boldsymbol{y}^{\mathbf{v}|a} - \boldsymbol{y}^{\mathbf{v}|b})(\boldsymbol{y}^o - \boldsymbol{y}^b)^{\mathrm{T}}\rangle\\ &\quad = \hat{\boldsymbol{P}}^b\{\mathbf{v}, o\}[\hat{P}^b + \boldsymbol{R}]^{-1}\langle(\boldsymbol{y}^o - \boldsymbol{y}^b)(\boldsymbol{y}^o - \boldsymbol{y}^b)^{\mathrm{T}}\rangle.\end{aligned} \tag{A6}$$

Then, Equation (22) is obtained from Equation (A3) by taking the transpose of both sides and substituting $\langle(\boldsymbol{y}^{\mathbf{v}|a} - \boldsymbol{y}^{\mathbf{v}|b})(\boldsymbol{y}^o - \boldsymbol{y}^b)^{\mathrm{T}}\rangle$ on the left-hand side (of the transpose of Equation (A3)) by the right-hand side of Equation (A6).

APPENDIX B. USING THE ENSEMBLE TO ESTIMATE COVARIANCE MATRICES

Kalman filter methods require error covariances of the model background state for which *ensemble* Kalman filters make the assumption that such covariances can be estimated from the respective ensemble covariances. Similarly, in such a framework, the analysis error covariance is related to the covariance of the analysis ensemble. Here, to compute analysis covariance $\hat{\boldsymbol{P}}^a\{\mathbf{v}, o\}$ (between verification and observation space) from the N_{ens} ensemble members of the LETKF, we define the incremental ensemble members in verification space $\mathbf{Y}^{\mathbf{v}|a}$ and observation space $\mathbf{Y}^a$ (with components $Y_\nu^{\mathbf{v}|a(i)}$ and $Y_\alpha^{a(i)}$, respectively). Let $y_\nu^{\mathbf{v}|a(i)}$ and $y_\alpha^{a(i)}$ be the values of $y_\nu^{\mathbf{v}|a}$ and y_α^a for the ith ensemble member; then, $Y_\nu^{\mathbf{v}|a(i)} = y_\nu^{\mathbf{v}|a(i)} - \overline{y_\nu^{\mathbf{v}|a}}$ is the difference from the corresponding ensemble mean value $\overline{y_\nu^{\mathbf{v}|a}}$ (and similarly $Y_\alpha^{a(i)} = y_\alpha^{a(i)} - \overline{y_\alpha^a}$).

With this we compute the ensemble covariance $\widetilde{P^b_{\mathrm{en}[\nu,\alpha]}}$ as

$$\widetilde{P^b_{\mathrm{en}[\nu,\alpha]}} = (N_{\mathrm{ens}} - 1)^{-1}\sum_{i=1}^{N_{\mathrm{ens}}}[Y_\nu^{\mathbf{v}|a(i)}Y_\alpha^{a(i)}] \tag{B1}$$

from which the estimator $\hat{P}^a_{\mathrm{en}[\nu,\alpha]}$ for the $[\nu, \alpha]$-component of $\hat{\boldsymbol{P}}^a\{\mathbf{v}, o\}$ is given by the product

$$\hat{P}^a_{\mathrm{en}[\nu,\alpha]} = \widetilde{P^b_{\mathrm{en}[\nu,\alpha]}} * \eta^t(\nu, \alpha), \tag{B2}$$

where $\eta^t(\nu, \alpha)$ is the localization function between the component $y_\nu^{\mathbf{v}|a}$ of the verifying data and the analysis observation y_α^a. (The corresponding equations for the background covariance matrix are obtained by replacing the superscript "a" by "b" in all the equations and definitions in this section.) We sometimes refer to $\hat{P}^a_{\mathrm{en}[\nu,\alpha]}$ as the localized ensemble covariance, whereas $\widetilde{P^a_{\mathrm{en}[\nu,\alpha]}}$ is the corresponding unlocalized covariance.

Localization is an essential technical procedure when estimating covariances via finite-size ensembles. It is needed to suppress spurious correlations that occur as a result of the small ensemble size. Most common is the localization in state space where $\eta(\mathbf{v}, \alpha)$ is a function of the distance between the two observations that goes to zero for large distances. Popular is the Gaspari–Cohn function (Gaspari and Cohn, 1999), which has some similarity with a Gaussian but has a finite support. Here, we use a superposition of Gaspari–Cohn functions gc(:) for the horizontal distance Δh and the difference in logarithmic pressure $\Delta[\log(p)]$ between the two observations:

$$\eta^t(\nu, \alpha) = \mathrm{gc}\left(\frac{\Delta h}{l_{\mathrm{h}}}\right)\mathrm{gc}\left(\frac{|\Delta[\log(p)]|}{l_{\mathrm{z}}}\right), \tag{B3}$$

where the vertical and horizontal localization length scales l_{z} and l_{h} are tunable input parameters.

APPENDIX C. STATISTICAL SIGNIFICANCE

When testing Equation (12), a central question is whether the sign of $S(J_\alpha^b)$ is statistically significant. All significance tests (e.g., like Student's t-test) have to make some assumption about the nature of the errors that might not be fulfilled for the data considered here. For the case where the amount of data is reasonably large (which is the case we are mostly interested in) these tests often give very similar results. Here, we have decided to use a significance indicator based on a very simple error model (and which therefore is very simple to apply) but which is sensitive to both the number and the magnitudes of the contributions in a respective bin.[7] More precisely, we compare the data A_i collected for a given bin with a model problem in which the data have the same magnitude but where the sign is completely random, which implies $\langle A_i\rangle = 0$ and $\langle A_iA_j\rangle = \delta_{ij}A_i^2$ so that also the sum over a bin

$$S(A) \equiv \sum_{i=1}^{m} A_i$$

has zero mean expectation value (i.e., $\langle S(A)\rangle = 0$) and the variance can be written as

$$\langle S(A)^2\rangle = \sum_{i=1}^{m}\sum_{j=1}^{m}\langle A_iA_j\rangle = \langle S(A^2)\rangle.$$

In the following, to get an impression whether the sign of the sum $S(\boldsymbol{J}_\alpha^b)$ computed for a given bin is statistically relevant, in the graphs presented in this article it is always compared with the standard deviation estimate $\mathbb{V}_\alpha^b$ of the corresponding noise model:

$$\mathbb{V}_\alpha^b = \sqrt{S[(J_\alpha^b)^2]}. \tag{C1}$$

Of course, since J_α^b is generally not a Gaussian variable (even for the case where the observational data are Gaussian), $\mathbb{V}_\alpha^b$ should only be seen as a rough indicator for which three standard deviations might generally not be sufficient to ensure statistical relevance.

APPENDIX D. SINGLE OBSERVATION DIAGNOSTICS

The diagnostics introduced in this article can also be applied to experiments where only a single observation (i.e., the observation α) is assimilated, in which case the corresponding analysis increments and analysis covariances are given by

$$\begin{aligned} y_\nu^{\mathbf{v}|a;\mathrm{SO}} - y_\nu^{\mathbf{v}|b} &= \hat{P}^b_{\mathrm{en}[\nu,\alpha]}\frac{y_\alpha^o - y_\alpha^b}{\hat{P}^b_{\alpha\alpha} + R_{\alpha\alpha}} \\ \hat{P}^{a;\mathrm{SO}}_{[\nu,\alpha]} &= \hat{P}^b_{\mathrm{en}[\nu,\alpha]}\frac{R_{\alpha\alpha}}{\hat{P}^b_{\alpha\alpha} + R_{\alpha\alpha}} \end{aligned} \tag{D1}$$

(where the additional superscript "SO" indicates that the respective quantity corresponds to single-observation assimilations). Introducing this into Equations 28a and 28b yields

$$J_\alpha^{b;\mathrm{SO}} = \sum_\nu \hat{P}^b_{\mathrm{en}[\nu,\alpha]}\frac{(y_\nu^{\mathbf{v}} - y_\nu^{\mathbf{v}|b})(y_\alpha^o - y_\alpha^b)}{R_{\nu\nu}(\hat{P}^b_{\alpha\alpha} + R_{\alpha\alpha})} \tag{D2a}$$

$$\begin{aligned} J_\alpha^{ab;\mathrm{SO}} &= \sum_\nu \hat{P}^b_{\mathrm{en}[\nu,\alpha]}\frac{(y_\nu^{\mathbf{v}|a;\mathrm{SO}} - y_\nu^{\mathbf{v}|b})(y_\alpha^o - y_\alpha^b)}{R_{\nu\nu}(\hat{P}^b_{\alpha\alpha} + R_{\alpha\alpha})} \\ &= \sum_\nu (\hat{P}^b_{\mathrm{en}[\nu,\alpha]})^2\frac{(y_\alpha^o - y_\alpha^b)^2}{R_{\nu\nu}(\hat{P}^b_{\alpha\alpha} + R_{\alpha\alpha})^2} \end{aligned} \tag{D2b}$$

while the expectation value, Equation (29), is

$$\langle J_\alpha^{b;\mathrm{SO}}\rangle_{\mathrm{estim}} = \sum_\nu \frac{(\hat{P}^b_{\mathrm{en}[\nu,\alpha]})^2}{R_{\nu\nu}(\hat{P}^b_{\alpha\alpha} + R_{\alpha\alpha})}. \tag{D3}$$

[7] Problems with statistical significance arise particularly when either the population in a bin is too small or when the sum is dominated by a small number of large contributions.