Author/s:
Carrio, DS;Bishop, CH;Kotsuki, S

Title:
Empirical determination of the covariance of forecast errors: An empirical justification and reformulation of hybrid covariance models

Date:
2021-04-01

Citation:
Carrio, D. S., Bishop, C. H. & Kotsuki, S. (2021). Empirical determination of the covariance of forecast errors: An empirical justification and reformulation of hybrid covariance models. Quarterly Journal of the Royal Meteorological Society, 147 (736), pp.2033-2052. https://doi.org/10.1002/qj.4008.

Persistent Link:
https://hdl.handle.net/11343/274394

CARRIÓ CARRIÓ DIEGO (Orcid ID: 0000-0002-4896-0794)
Kotsuki Shunji (Orcid ID: 0000-0002-0717-9192)

Empirical determination of the covariance of forecast errors: an empirical justification and reformulation of Hybrid covariance models.

D. S. Carrió[a,b], C. H. Bishop[a,b], S. Kotsuki[c,d,e]

[a]*School of Earth Sciences. The University of Melbourne, Parkville, Victoria, Australia*

[b]*ARC Centre of Excellence for Climate Extremes*

[c]*Center for Environmental Remote Sensing, Chiba, Kobe, Japan*

[d]*RIKEN Center for Computation*

[e]*PRESTO, Japan Science and Technology Agency, Chiba, Japan*

ABSTRACT

During the last decade, the replacement of static climatological forecast error covariance models with Hybrid error covariance models, that linearly combine localized ensemble covariances with static climatological error covariances, has led to significant forecast improvements at several major forecasting centres. Here, a deeper understanding of why the Hybrid's superficially *ad hoc* mix of ensemble-based and climatological covariances yield such significant improvements is pursued. In practice, ensemble covariances are not equal to the true flow-dependent forecast error covariance matrix. Here, the relationship between actual forecast error covariance and the corresponding ensemble covariance is empirically demonstrated. Using a simplified Global Circulation Model and the Local Ensemble Transform Kalman Filter (LETKF), the covariance of the set of actual forecast errors corresponding to ensemble covariances close to a fixed target value is computed. By doing this for differing target values, an estimate of the actual forecast error covariance as a function of ensemble covariance is obtained. A demonstration that the Hybrid is a much better approximation to this

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/qj.4008

estimate than either the static climatological covariance or the localized ensemble covariance is given. The empirical estimate has two features that current Hybrid error covariance models fail to represent: (i) The weight given to the static covariance matrix is an increasing function of the horizontal separation distance of the covarying model variables, and (ii) For small ensemble sizes and ensemble covariances near zero but negative, the actual forecast error covariance is a decreasing function of increasing ensemble covariance. While the first finding has been anticipated by other authors, the second finding has not been anticipated, as far as the authors are aware. Here, (ii) is hypothesized to be consequence of spurious sample correlations and variances associated with reduced ensembles. Consistent with this hypothesis, the non-monotonicity of this relationship is almost eliminated by quadrupling the ensemble size.

Correspondence: D. S. Carrió, ARC Centre of Excellence for Climate Extremes and School of Earth Sciences, The University of Melbourne, Parkville 3010, Victoria, Australia

Email: diego.carriocarrio@unimelb.edu.au

Funding Information: ARC Centre of Excellence for Climate Extremes (CE170100023)

Keywords: Data assimilation, ETKF, Hybrid forecast error covariance model, ensemble-based covariance matrix, climatological error covariance matrix, R-localisation function, B-localisation function

1. Introduction

A key component of state-of-the-art Data Assimilation (DA) schemes used to obtain initial conditions for weather forecasts is the background or forecast error covariance matrix**.** A number of studies have investigated the effect on DA performance of replacing a quasi-static climatological estimate of the forecast error covariance by a Hybrid forecast error covariance model (e.g., Clayton et al., (2013); Kuhl et al., (2013); Wang et al., (2014)). The Hybrid covariance model linearly combines an ensemble covariance [1] matrix with a static climatological covariance matrix. These studies have shown that switching to the Hybrid covariance model leads to significant forecast improvements. A possible reason for these improvements is that the Hybrid model of the forecast error covariance matrix is closer to the covariance matrix of true forecast errors given known antecedents, also known as true Bayesian forecast error covariance matrix. Theory states that DA schemes such as 4D-Var DA and the EnKF maximise their potential accuracy when the value of the forecast error covariance matrix used in the DA scheme is equal to the covariance matrix of the relevant distribution of forecast errors (e.g., Daley (1993); Courtier et al., (1994); Kalnay (2003)). This ideal covariance matrix is highly flow dependent because it depends on error growth dynamics linked to the actual state and also on the location and accuracy of observations assimilated at previous DA cycles. If computational resources were unlimited and accurate descriptions of model error were available, Bayes' theorem could be solved using methods such as the Particle Filter (e.g., Snyder et al., (2008); Van Leeuwen (2010); Van Leeuwen (2012); Reich et al., (2015); Van Leeuwen (2019)) to accurately estimate the distribution of flow dependent forecast errors and their covariances. However, in practice, all that is available is an ensemble forecast with order

[1] Note that the term "ensemble covariance" is a widely used term in DA and it refers to as an ensemble-based approximation to the true forecast error covariance matrix.

10-100 members and some *ad hoc* representation of model error together with a suboptimal solution to Bayes' equations such as the Ensemble Kalman Filter (EnKF; e.g., Houtekamer and Mitchell, (1998); Evensen (1994); Anderson (2001); Bishop et al., (2001); Whitaker and Hamill, (2002); Houtekamer and Mitchell, (2005); Hamill (2006); Bonavita et al., (2008)). Consequently, neither the ensemble covariance matrix $\mathbf{P}_{ens}^{f}$ nor the localized ensemble covariance matrix $\mathbf{P}_{ens}^{f} \circ \mathbf{C}$ give the true flow dependent forecast error covariance matrix. They are imperfect. The idea that a linear combination of an estimate $\mathbf{B}$ of the static climatological error covariance and $\mathbf{P}_{ens}^{f} \circ \mathbf{C}$ might give a better approximation to the true forecast error covariance matrix, was first suggested by Hamill et al., (2000) (as far as the authors are aware).

For the case of an idealized model, Bishop and Satterfield, (2013) used a replicate system approach to empirically identify (i) the distribution of all true forecast error variances experienced by the DA forecasting system over a large number of cycles (ii) the distribution of imperfect ensemble sample variances given a true forecast error variance, and (iii) the distribution of true forecast error variances given an ensemble variance. They noted that the key quantity required for DA was the mean of the (*unknown*) true forecast error variance distribution as a function of the *known* imperfect ensemble variance. They provided both empirical and theoretical support for the notion that the covariance of true forecast errors given an ensemble variance would be approximately equal to a Hybrid linear combination of the climatological error variance of a climatology of forecast errors and the flow-dependent ensemble variance. However, Bishop and Satterfield's (2013) results were confined to *variances* and did not consider covariances. This paper extends the work of Bishop and Satterfield, (2013) from forecast error variances to forecast error covariances. It does so by empirically describing the mean of the distribution of true forecast error covariances given (i)

the corresponding ensemble covariance, and (ii) the separation distance[2] between the two variables whose error covariance is being examined. It can be shown that the mean of the distribution of the true error covariances given a specific ensemble covariance value is equal to the covariance of the set of forecast errors corresponding to ensemble covariances whose values happen to be very close to the specified ensemble covariance value. Note that, ideally, ensembles and their covariances capture and quantify important aspects of the flow dependent error covariance that one might qualitatively expect from an examination of the ensemble mean. For example, if an intense cyclone is present in the ensemble mean, differing ensemble members would likely contain cyclones of differing intensities at differing locations and the covariance of the ensemble would thus reflect the type of flow dependent error covariances one would expect in this situation. Thus, ideally, an ensemble captures all of the error covariance information one might deduce from knowledge of the ensemble mean. We shall refer to this conditional mean of the (unknown) distribution of true forecast error covariances as the actual covariance of forecast error covariances given an ensemble covariance. As will be seen, the results give insight into the meaning and necessity of ensemble covariance localisation and how ensemble covariance localisation should, in fact, be directly linked to the relative weight given to elements of the climatological forecast error covariance matrix.

Other notable work in the area of Hybrid error covariance modelling and ensemble covariance localization includes that of Ménétrier and Auligné (2015). They derive optimal ensemble covariance localization functions and a diagonal weight matrix for the static climatological covariance matrix based on the assumption that the ensemble samples the true Bayesian flow dependent distribution of forecast errors given past observations. Here, we explicitly recognize that the ensemble generated by a 20-80 member practicable ensemble DA

[2] In general, the separation distance would be defined by a suitable metric in 3D space. Here, we shall only consider variables at the same model level and separation distance is simply the great circle horizontal distance between these two points.

scheme is *not* drawn from the true flow dependent distribution. In addition, we do not confine ourselves to the consideration of Hybrid covariance models where the weight on the static covariance model is defined by diagonal matrices. In contrast, we empirically examine how the covariance of actual forecast errors varies with changes in imperfect ensemble covariances. Having done that, we then assess how well previously proposed Hybrid covariance models account for the empirically defined relationships. Section 2 gives the general form of the minimum error variance state estimation equations *given an imperfect ensemble covariance matrix*. In Section 3, a general method for computing the covariance of forecast errors as a function of ensemble covariance and variable separation distance is given. Section 4 introduces the numerical weather model used to perform our experiments and describes the experimental setup followed in this study. In Section 5, results obtained from our numerical experiments are presented, discussed and interpreted. Concluding remarks follow in section 6. A supplementary table (Table S1) listing all the symbols and their meaning is also included.

2. The linear state estimation equations that minimize analysis error variance given an ensemble covariance

This section serves to (i) prove that the covariances of actual forecast errors given the corresponding covariances of an imperfect ensemble are required by the Kalman gain matrix to minimize analysis error variance, (ii) develop the equations required to analyse DA performance in the presence of an inaccurate forecast error covariance model, and (iii) introduce our mathematical notation.

Let $\mathbf{x}$ be a vector that lists all of the variables defining a model state (see Table 1 for a list of symbols). Given (i) the mean $\overline{\mathbf{x}}^f$ of the prior distribution of possible states, (ii) a list of imperfect observations $\mathbf{y} = H(\mathbf{x}^t) + \varepsilon^o$, where H is the nonlinear observation operator, $\mathbf{x}^t$ is

the true model state and ε^o is the observation error, and (iii) the mean of the prior pdf of the observed variable $\overline{H\left(\mathbf{x}^f\right)}$, the equation for the posterior mean $\bar{\mathbf{x}}^a$ for the EnKF (and also 4D-Var in the linear case) takes the form:

$$\begin{aligned}\bar{\mathbf{x}}^a &= \bar{\mathbf{x}}^f + \mathbf{K}\left(\mathbf{y} - \overline{H\left(\mathbf{x}^f\right)}\right) \\ &= \bar{\mathbf{x}}^f + \mathbf{K}\left(H\left(\mathbf{x}^t\right) + \varepsilon^o - \overline{H\left(\mathbf{x}^f\right)}\right), \text{ where } H\left(\mathbf{x}^t\right) \text{ is the true value of the observed variable} \\ &= \bar{\mathbf{x}}^f + \mathbf{K}\left(\varepsilon^o - \varepsilon_o^f\right), \text{ where } \varepsilon_o^f = \overline{H\left(\mathbf{x}^f\right)} - H\left(\mathbf{x}^t\right) \text{ is the forecast error in observation space.}\end{aligned} \quad (1)$$

Subtracting the unknown true state $\mathbf{x}^t$ from both sides of Equation 1 gives:

$$\begin{aligned}&\bar{\mathbf{x}}^a - \mathbf{x}^t = \bar{\mathbf{x}}^f - \mathbf{x}^t + \mathbf{K}\left(\varepsilon^o - \varepsilon_o^f\right) \\ &\Rightarrow \varepsilon^a = \varepsilon^f + \mathbf{K}\left(\varepsilon^o - \varepsilon_o^f\right), \text{ where } \varepsilon^a = \bar{\mathbf{x}}^a - \mathbf{x}^t \text{ and } \varepsilon^f = \bar{\mathbf{x}}^f - \mathbf{x}^t\end{aligned} \quad (2)$$

Now let ε_i^a and ε_i^f denote the i^{th} element of the analysis and forecast error vectors, respectively. Similarly, let $\mathbf{K}_i$ denote the i^{th} row of the *gain matrix* **K.**

Equation 2 then implies that the product of the errors in the analysis of the i^{th} and j^{th} state variables is given by:

$$\begin{aligned}\varepsilon_i^a\left(\varepsilon_j^a\right)^T &= \left[\varepsilon_i^f + \mathbf{K}_i\left(\varepsilon^o - \varepsilon_o^f\right)\right]\left[\varepsilon_j^f + \mathbf{K}_j\left(\varepsilon^o - \varepsilon_o^f\right)\right]^T \\ &= \left[\varepsilon_i^f + \mathbf{K}_i\left(\varepsilon^o - \varepsilon_o^f\right)\right]\left[\left(\varepsilon_j^f\right)^T + \left(\varepsilon^o - \varepsilon_o^f\right)^T \mathbf{K}_j^T\right] \\ &= \varepsilon_i^f\left(\varepsilon_j^f\right)^T + \varepsilon_i^f\left(\varepsilon^o - \varepsilon_o^f\right)^T \mathbf{K}_j^T + \mathbf{K}_i\left(\varepsilon^o - \varepsilon_o^f\right)\left(\varepsilon_j^f\right)^T + \mathbf{K}_i\left(\varepsilon^o - \varepsilon_o^f\right)\left(\varepsilon^o - \varepsilon_o^f\right)^T \mathbf{K}_j^T\end{aligned} \quad (3)$$

Note that here, ε_j^f is a scalar and hence $\varepsilon_j^{fT} = \varepsilon_j^f$. Hence, we can rewrite Equation 3 as:

$$\varepsilon_i^a\varepsilon_j^a = \varepsilon_i^f\varepsilon_j^f + \varepsilon_i^f\left(\varepsilon^o - \varepsilon_o^f\right)^T \mathbf{K}_j^T + \mathbf{K}_i\left(\varepsilon^o - \varepsilon_o^f\right)\varepsilon_j^f + \mathbf{K}_i\left(\varepsilon^o - \varepsilon_o^f\right)\left(\varepsilon^o - \varepsilon_o^f\right)^T \mathbf{K}_j^T \quad (4)$$

The next step in our derivation is to take the expected value of Equation 4 over some distribution of errors. If the ensemble had an infinite number of members and was perfectly designed, then these ensemble covariances would be precisely equal to the fully *flow and observation dependent* covariance of the infinite set of forecast errors that could occur given all antecedent observations. In practice, ensemble forecasting schemes fail to accurately account for all sources of uncertainty, and hence are inaccurate. Nevertheless, one expects some of the covariance of actual forecast errors to be stochastically linked to the ensemble covariance and vice-versa.

If forecast errors were accurately known, one could compute the rank-1 outer product matrix $\boldsymbol{\varepsilon}^f \boldsymbol{\varepsilon}^{fT}$ corresponding to each full ensemble covariance matrix $\mathbf{P}^f_{ens}$. Cycling the DA scheme for an infinite number of cycles would then yield an infinite number of $\left(\boldsymbol{\varepsilon}^f \boldsymbol{\varepsilon}^{fT}, \mathbf{P}^f_{ens}\right)$ pairs. Given this infinite set, one could then collect a very large subset of these pairs all having very similar values of the $\mathbf{P}^f_{ens}$ matrix (see Section 3 for further details on how we obtain this subset in practice). The average $\left\langle \boldsymbol{\varepsilon}^f \boldsymbol{\varepsilon}^{fT} \mid \mathbf{P}^f_{ens} \right\rangle$ of the pairs $\left(\boldsymbol{\varepsilon}^f \boldsymbol{\varepsilon}^{fT}, \mathbf{P}^f_{ens}\right)$ would then be the covariance of forecast errors that occur on those occasions where $\mathbf{P}^f_{ens}$ has essentially the *same* value. We denote this forecast error covariance matrix by $\mathbf{P}^f = \left\langle \boldsymbol{\varepsilon}^f \boldsymbol{\varepsilon}^{fT} \mid \mathbf{P}^f_{ens} \right\rangle$ and the scalar element of $\mathbf{P}^f$ lying on the i^{th} row and j^{th} column by P^f_{ij}. See Appendix for a theoretical justification that $\mathbf{P}^f$ can be approximated as the average of the forecast errors $\boldsymbol{\varepsilon}^f \boldsymbol{\varepsilon}^{fT}$ over its climatological distribution conditioned on $\mathbf{P}^f_{ens}$. In the same way, we can define $\mathbf{P}^f H^T = \left\langle \boldsymbol{\varepsilon}^f \left(\boldsymbol{\varepsilon}^f_o\right)^T \mid \mathbf{P}^f_{ens} H^T \right\rangle$, where $\mathbf{P}^f_{ens} H^T \equiv \frac{1}{N-1} \sum_{i=1}^{N} \left(\mathbf{x}^f_i - \overline{\mathbf{x}^f}\right)\left(H\left(\mathbf{x}^f_i\right) - \overline{H\left(\mathbf{x}^f\right)}\right)^T$ and $H\mathbf{P}^f H^T = \left\langle \left(\boldsymbol{\varepsilon}^f_o\right)\left(\boldsymbol{\varepsilon}^f_o\right)^T \mid H\mathbf{P}^f_{ens} H^T \right\rangle$, where $H\mathbf{P}^f_{ens} H^T \equiv \frac{1}{N-1} \sum_{i=1}^{N} \left(H\left(\mathbf{x}^f_i\right) - \overline{H\left(\mathbf{x}^f\right)}\right)\left(H\left(\mathbf{x}^f_i\right) - \overline{H\left(\mathbf{x}^f\right)}\right)^T$,

where N is the number of ensemble members. In the above equations, we have followed Houtekamer and Mitchell (2001) and used the symbols $\mathbf{P}^f H^T$, $\mathbf{P}^f_{ens} H^T$, $H\mathbf{P}^f H^T$ and $H\mathbf{P}^f_{ens} H^T$ to remind the reader that the covariances between model and observation space referred to by these symbols were computed using the full possibly *nonlinear* observation operator H and not its (purely linear) Jacobian $\mathbf{H}$. Hence, the reader is cautioned to *not* interpret the symbols $\mathbf{P}^f H^T$, $\mathbf{P}^f_{ens} H^T$, $H\mathbf{P}^f H^T$ and $H\mathbf{P}^f_{ens} H^T$ as meaning that we have confined our analysis to cases where the observation operator is linear. With this notation, the expected value of Equation 4 *given* the ensemble covariance matrix then yields:

$$\begin{aligned}
\left\langle \varepsilon_i^a \varepsilon_j^a \right\rangle &= \left\langle \varepsilon_i^f \varepsilon_j^f \right\rangle + \left\langle \varepsilon_i^f \left(\varepsilon^o - \varepsilon_o^f\right)^T \mathbf{K}_j^T \right\rangle + \left\langle \mathbf{K}_i \left(\varepsilon^o - \varepsilon_o^f\right) \varepsilon_j^f \right\rangle + \left\langle \mathbf{K}_i \left(\varepsilon^o - \varepsilon_o^f\right)\left(\varepsilon^o - \varepsilon_o^f\right)^T \mathbf{K}_j^T \right\rangle \\
&= \left\langle \varepsilon_i^f \varepsilon_j^f \right\rangle + \left\langle \varepsilon_i^f \left(\varepsilon^o - \varepsilon_o^f\right)^T \right\rangle \mathbf{K}_j^T + \mathbf{K}_i \left\langle \left(\varepsilon^o - \varepsilon_o^f\right) \varepsilon_j^f \right\rangle + \mathbf{K}_i \left\langle \left(\varepsilon^o - \varepsilon_o^f\right)\left(\varepsilon^o - \varepsilon_o^f\right)^T \right\rangle \mathbf{K}_j^T \\
&= \left\langle \varepsilon_i^f \varepsilon_j^f \right\rangle + \left\langle \varepsilon_i^f \left(\varepsilon^o\right)^T \right\rangle \mathbf{K}_j^T - \left\langle \varepsilon_i^f \left(\varepsilon_o^f\right)^T \right\rangle \mathbf{K}_j^T + \mathbf{K}_i \left\langle \varepsilon^o \varepsilon_j^f \right\rangle - \mathbf{K}_i \left\langle \varepsilon_o^f \varepsilon_j^f \right\rangle + \mathbf{K}_i \left\langle \varepsilon^o \left(\varepsilon^o\right)^T \right\rangle \mathbf{K}_j^T \\
&\quad - \mathbf{K}_i \left\langle \varepsilon^o \left(\varepsilon_o^f\right)^T \right\rangle \mathbf{K}_j^T - \mathbf{K}_i \left\langle \varepsilon_o^f \left(\varepsilon^o\right)^T \right\rangle \mathbf{K}_j^T + \mathbf{K}_i \left\langle \varepsilon_o^f \left(\varepsilon_o^f\right)^T \right\rangle \mathbf{K}_j^T \\
&= P_{ij}^f - \mathbf{P}_i^f H^T \mathbf{K}_j^T - \mathbf{K}_i \left(\mathbf{P}_j^f H^T\right)^T + \mathbf{K}_i \mathbf{R} \mathbf{K}_j^T + \mathbf{K}_i \left(H\mathbf{P}^f H^T\right) \mathbf{K}_j^T \\
&= P_{ij}^f - \mathbf{P}_i^f H^T \mathbf{K}_j^T - \mathbf{K}_i \left(\mathbf{P}_j^f H^T\right)^T + \mathbf{K}_i \left(H\mathbf{P}^f H^T + \mathbf{R}\right) \mathbf{K}_j^T
\end{aligned} \tag{5}$$

$$P_{ij}^a = P_{ij}^f - \mathbf{P}_i^f H^T \left(\mathbf{K}_j\right)^T - \mathbf{K}_i \left(\mathbf{P}_j^f H^T\right)^T + \mathbf{K}_i \left(H\mathbf{P}^f H^T + \mathbf{R}\right) \left(\mathbf{K}_j\right)^T \tag{6}$$

where $\mathbf{P_i}$ and $\mathbf{K}_i$ denote the i^{th} row of the matrices $\mathbf{P}^f$ and $\mathbf{K}$, respectively. Note that here we have assumed that the observational and forecast errors are not correlated $\left\langle \varepsilon^o \varepsilon_j^f \right\rangle = \left\langle \varepsilon_j^f \left(\varepsilon^o\right)^T \right\rangle = \mathbf{0}$, which is a typical assumption in the derivation of Kalman filter DA algorithms. In addition, we have also defined the observational error covariance matrix as $\left\langle \varepsilon^o \left(\varepsilon^o\right)^T \right\rangle = \mathbf{R}$. In the case that i=j, Equation 6 reduces to the equation

$$P_{\bar{i}\bar{i}}^{a} = P_{\bar{i}\bar{i}}^{f} - \mathbf{P}_i^f H^T \left(\mathbf{K}_i\right)^T - \mathbf{K}_i \left(\mathbf{P}_i^f H^T\right)^T + \mathbf{K}_i \left(H\mathbf{P}^f H^T + \mathbf{R}\right)\left(\mathbf{K}_i\right)^T$$
$$= P_{\bar{i}\bar{i}}^{f} - 2\mathbf{K}_i \left(\mathbf{P}_i^f H^T\right)^T + \mathbf{K}_i \left(H\mathbf{P}^f H^T + \mathbf{R}\right)\left(\mathbf{K}_i\right)^T \quad (7)$$

Note that Equation 7 gives the analysis error variance $P_{\bar{i}\bar{i}}^{a}$ as a quadratic function of the elements of the row vector $\mathbf{K}_i$. It follows that the choice of $\mathbf{K}_i$ that minimizes $P_{\bar{i}\bar{i}}^{a}$ is the choice that makes all of the elements of the derivative vector $\frac{\partial P_{\bar{i}\bar{i}}^{a}}{\partial\left(\mathbf{K}_i\right)^T}$ equal to zero, where the i^{th} element of the column vector $\frac{\partial P_{\bar{i}\bar{i}}^{a}}{\partial\left(\mathbf{K}_i\right)^T}$ lists the derivative of P_{ii}^{a} with respect to the i^{th} element of the row vector $\mathbf{K}_i$. Taking the derivative of Equation 7 gives

$$\frac{\partial P_{\bar{i}\bar{i}}^{a}}{\partial \mathbf{K}_i^T} = -2\left(\mathbf{P}_i^f H^T\right)^T + 2\left(H\mathbf{P}^f H^T + \mathbf{R}\right)\mathbf{K}_i^T \quad (8)$$

Setting this derivative to zero then gives the i^{th} row of the optimal gain matrix given the ensemble covariance matrix as,

$$\mathbf{K}_i^T = \left(H\mathbf{P}^f H^T + \mathbf{R}\right)^{-1}\left(\mathbf{P}_i^f H^T\right)^T$$
$$\Rightarrow \mathbf{K}_i = \mathbf{P}_i^f H^T \left(H\mathbf{P}^f H^T + \mathbf{R}\right)^{-1} \quad (9)$$

Equation 9 relates the row vector $\mathbf{K}_i$ to the i^{th} row of $\mathbf{P}_i^f H^T$; hence, the full gain matrix is simply

$$\mathbf{K} = \mathbf{P}^f H^T \left(H\mathbf{P}^f H^T + \mathbf{R}\right)^{-1} \quad (10)$$

Note that the derivation of Equation 10 shows that the form of the gain matrix which minimizes the analysis error variance (i.e., optimal gain) is given in terms of the covariance of actual

forecast errors given specific ensemble covariance values of $\mathbf{P}^f_{ens}$ and the observational error covariance matrix $\mathbf{R}$.

It can be shown that the number of trials $\left(\varepsilon^f \varepsilon^{fT}, \mathbf{P}^f_{ens}\right)$ that would be required to empirically estimate $\mathbf{P}^f = \left\langle \varepsilon^f \varepsilon^{fT} \,|\, \mathbf{P}^f_{ens} \right\rangle$ for a cycling ensemble DA scheme is prohibitively large. Nevertheless, there are good reasons to believe that good approximations to $\mathbf{P}^f = \left\langle \varepsilon^f \varepsilon^{fT} \,|\, \mathbf{P}^f_{ens} \right\rangle$ can be obtained with smaller data sets. To begin, note that in the limit of a perfect ensemble forecast, the *ij*th element $\left\{\mathbf{P}^f_{ens}\right\}_{ij}$ of $\mathbf{P}^f_{ens}$ is precisely equal to the *ij*th element $\left\{\mathbf{P}^f\right\}_{ij} = \left\langle \varepsilon^f_i \varepsilon^f_j \,|\, \mathbf{P}^f_{ens} \right\rangle$ of $\mathbf{P}^f = \left\langle \varepsilon^f \varepsilon^{fT} \,|\, \mathbf{P}^f_{ens} \right\rangle$. Furthermore, for a fairly well constructed ensemble, one would expect changes in the element $\left\{\mathbf{P}^f_{ens}\right\}_{ij}$ to be the primary predictor of changes in $\left\{\mathbf{P}^f\right\}_{ij} = \left\langle \varepsilon^f_i \varepsilon^f_j \,|\, \mathbf{P}^f_{ens} \right\rangle$. Hence, here we examine the approximation to $\left\{\mathbf{P}^f\right\}_{ij} = \left\langle \varepsilon^f_i \varepsilon^f_j \,|\, \mathbf{P}^f_{ens} \right\rangle$ given by $\left\{\mathbf{P}^f\right\}_{ij} = \left\langle \varepsilon^f_i \varepsilon^f_j \,|\, \left\{\mathbf{P}^f_{ens}\right\}_{ij} \right\rangle$ - which is much easier to estimate empirically. As in the Hollingsworth-Lönnberg method (Hollingsworth and Lönnberg, 1986), the full matrix $\mathbf{P}^f$ constructed from individual estimates of $\left\{\mathbf{P}^f\right\}_{ij} = \left\langle \varepsilon^f_i \varepsilon^f_j \,|\, \left\{\mathbf{P}^f_{ens}\right\}_{ij} \right\rangle$ is not guaranteed to be positive definite – even though it would be if the ensemble covariances were perfectly accurate and the ensemble sample size was infinite. Nevertheless, there are approaches for making small adjustments to the approximate matrix to remove such deficiencies such as removing the eigenvectors corresponding to negative and zero eigen-values from this matrix. If this positive definite matrix obtained from accurate empirical estimates of $\left\{\mathbf{P}^f\right\}_{ij} = \left\langle \varepsilon^f_i \varepsilon^f_j \,|\, \left\{\mathbf{P}^f_{ens}\right\}_{ij} \right\rangle$ were used in the gain matrix given by Equation 9 to make analyses, these analyses would be close to the minimum error variance estimates for the

system and hence they would likely be more accurate than those obtained from, say, a DA scheme that had, for example, incorrectly assumed that $\mathbf{P}^f = \mathbf{P}^f_{ens}$. Amongst other things, we will empirically demonstrate that Hybrid error covariance models can give covariances much closer to $\left\{\mathbf{P}^f\right\}_{ij} = \left\langle \varepsilon^f_i \varepsilon^f_j \,|\, \left\{\mathbf{P}^f_{ens}\right\}_{ij} \right\rangle$ than either a pure ensemble covariance model or a climatological model of error covariance.

3. General method for computing the covariance of forecast errors that occur for similar ensemble covariance values as a function of the ensemble covariance values and also of the variable separation distance

(For reference, note that subsection 4.2 gives the specific method we used in our study.)

In a cycling ensemble forecasting system, analyses are made every δt hours using the forecast from the previous analysis plus observations that have been collected over the antecedent δt hours. Ideally, the DA procedure estimates its own uncertainty with an ensemble of analyses from which a subsequent ensemble forecast is launched, and this ensemble of forecasts is used in creating the next analysis, and so on. Here we assume that the number, locations and accuracy of the observations used in the cycled DA scheme are similar from one cycle to the next or at least exhibit some periodicity over some number of DA cycles. For simplicity, here we confine our consideration to atmospherically relevant idealised DA systems in which the truth is known. Hereafter, and in order to simplify the notation in the further equations, we denote $\left\{\mathbf{P}^f_{ens}\right\}_{ij}$ as P^{ens}_{ij}. Now, let $P^f_{ij}\left(P^{ens}_{ij}, d_{ij}\right)$ denote the covariance of forecast errors of the i^{th} and j^{th} model variable as a function of the corresponding covariance of the ensemble forecast P^{ens}_{ij} and the separation distance d_{ij} between the i^{th} and j^{th} model variable. Also, let $B^c_{ij}\left(d_{ij}\right)$ denote the static climatological forecast error covariance between forecast

errors of the i^{th} and j^{th} model variable as a function solely of the separation distance d_{ij} regardless of the value of P_{ij}^{ens}.

To reiterate, $B_{ij}^{c}\left(d_{ij}\right)$ is the covariance over the *entire* archive of forecast errors of the forecast error between model variables that are separated by the distance $d_{ij} = b$, where b is a fixed scalar constant. In contrast, if $P_{ij}^{ens} \approx a$ and $d_{ij} \approx b$, then $P_{ij}^{f}\left(P_{ij}^{ens}, d_{ij}\right)$ is the covariance of the *subset* of the archive of forecast errors for which both P_{ij}^{ens} and d_{ij} happen to be approximately equal to the fixed constant scalars a and b, respectively. Hence, $B_{ij}^{c}\left(d_{ij}\right)$ is identical to the average of $P_{ij}^{f}\left(P_{ij}^{ens}, d_{ij}\right)$ over all occurrences of P_{ij}^{ens}.

The function $P_{ij}^{f}\left(P_{ij}^{ens}, d_{ij}\right)$ can be accurately estimated by performing a very large number of iterations (i.e., n =1, 2, …, N, where $N \to \infty$) of the DA forecasting system while performing the following computations:

a. For the n^{th} forecast, compute the forecast error for each model variable of interest. Store these variables in the error vector ε^{n}.

b. Compute the matrix $\mathbf{E}^{n} = \varepsilon^{n}\varepsilon^{nT}$. Note that each element of this matrix is simply the product of the contemporaneous forecast errors of two model variables. Denote the corresponding elements on the i^{th} row and j^{th} column of $\mathbf{E}^{n}$ and $\left(\mathbf{P}_{ens}^{f}\right)^{n}$ by e_{ij}^{n} and $\left(P_{ij}^{ens}\right)^{n}$, respectively.

c. Compute the distance d_{ij}^{n} between the variables associated with e_{ij}^{n} and $\left(P_{ij}^{ens}\right)^{n}$.

These computations produce data triplets $\left[e_{ij}^{n},\left(P_{ij}^{ens}\right)^{n},d_{ij}^{n}\right]$. These data triplets can be used to evaluate the actual forecast error covariance $P_{ij}^{f}\left[\left(P_{ij}^{ens}\right)_{\text{target}},\left(d_{ij}\right)_{\text{target}}\right]$ at target values $\left(P_{ij}^{ens}\right)_{\text{target}}$ **and** $\left(d_{ij}\right)_{\text{target}}$. This can be done by (i) ordering the triplets $\left[e_{ij}^{n},\left(P_{ij}^{ens}\right)^{n},d_{ij}^{n}\right]$ from smallest separation distance to largest separation distance (ii) putting the triplets into approximately equally populated bins based on separation distance (the mean value of d_{ij} within each bin then becomes the $\left(d_{ij}\right)_{\text{target}}$ associated with that bin) (iii) ordering all of the triplets within each separation distance bin from smallest to largest value of P_{ij}^{ens} and then dividing them into approximately equally populated bins with similar values of P_{ij}^{ens} (the mean value of P_{ij}^{ens} within each of these bins then becomes the $\left(P_{ij}^{ens}\right)_{\text{target}}$ value for that bin). Note that steps (ii) and (iii) then yield bins of data triplets $\left[e_{ij}^{n},\left(P_{ij}^{ens}\right)^{n},d_{ij}^{n}\right]$ each associated with a unique $\left(P_{ij}^{ens}\right)_{\text{target}}$ **and** $\left(d_{ij}\right)_{\text{target}}$ combination. The function $P_{ij}^{f}\left[\left(P_{ij}^{ens}\right)_{\text{target}},\left(d_{ij}\right)_{\text{target}}\right]$ is then approximated by the average value of all the e_{ij}^{n} values within each bin. We denote this approximation by $\overline{P_{ij}^{f}\left[\left(P_{ij}^{ens}\right)_{\text{target}},\left(d_{ij}\right)_{\text{target}}\right]}$. The function $B_{ij}^{c}\left[\left(d_{ij}\right)_{\text{target}}\right]$ is then approximated by the average value of all the e_{ij}^{n} values within all bins having the same $\left(d_{ij}\right)_{\text{target}}$ value, *regardless of the value of* $\left(P_{ij}^{ens}\right)_{\text{target}}$ within the bin. We denote this approximation by $\overline{B_{ij}^{c}\left[\left(d_{ij}\right)_{\text{target}}\right]}$. Note that $\overline{B_{ij}^{c}\left[\left(d_{ij}\right)_{\text{target}}\right]}$ is simply the average of $\overline{P_{ij}^{f}\left[\left(P_{ij}^{ens}\right)_{\text{target}},\left(d_{ij}\right)_{\text{target}}\right]}$ values having the same value of $\left(d_{ij}\right)_{\text{target}}$ but

differing values of $\left(P_{ij}^{ens}\right)_{\text{target}}$. In order for this approximation to be accurate, one must have performed sufficient DA cycles N to ensure that there are enough e_{ij}^{n} values within each bin to ensure that the computed average is statistically stable. In determining an acceptable sample size, it is helpful to remember that in the case that the errors are normally distributed, the variance of the sample covariance $\overline{P_{ij}^{f}\left[\left(P_{ij}^{ens}\right)_{\text{target}},\left(d_{ij}\right)_{\text{target}}\right]}$ is proportional to $\frac{1}{N-1}$ (Wishart (1928); Press (2012)). Hence, every time we double the sample size, the variance of stochastic variations in our estimation associated with finite sample size is approximately halved. With this fact in mind, in Section 4, we deliberately halve our sample size to see if our covariance estimates are changed by reducing the sample size. We find that our estimates are insensitive to whether we use the first or last half of our data set. This result is consistent with our hypothesis that we did perform sufficient DA cycles to ensure statistically stable covariance estimates.

4. Illustration using SPEEDY-LETKF Global DA Model

Here, we illustrate the above estimation method and compare $\overline{P_{ij}^{f}\left[\left(P_{ij}^{ens}\right)_{\text{target}},\left(d_{ij}\right)_{\text{target}}\right]}$ to other commonly used forecast error covariance models (including the Hybrid).

4.1 The SPEEDY-LETKF system

To simply represent flow structures and initial condition sensitivity similar to that found in the atmosphere, we employ a simplified, primitive equation based global atmospheric model called SPEEDY (Molteni, 2003). The SPEEDY model is a hydrostatic, spectral-transform

model in the vorticity-divergence form, with a semi-implicit treatment of gravity waves. The equivalent grid dimension used are 96x48 grid points in the horizontal (T30 ~3.75ºx3.75º) and 7 vertical layers using the sigma-coordinate system. Zonal physical distances between grid points are approximately 417 km on the equator, 361 km at the tropics (i.e., at 23N and 23S) and 208 km at 60N and 60S. There are 7 sigma (pressure) vertical levels, namely 0.950 (925 hPa), 0.835 (850 hPa), 0.685 (700 hPa), 0.510 (500 hPa), 0.340 (300 hPa), 0.200 (200 hPa) and 0.080 (100 hPa). The SPEEDY dynamical state variables are temperature, specific humidity, zonal wind, and meridional wind and surface pressure. While running the SPEEDY model is computationally inexpensive, it contains physical parameterizations for cloud, condensation, convection, radiation and surface-fluxes.

In order to obtain a "truth" run as well as error prone observations for our idealized experiments, we first performed a 1-year simulation (without applying DA) in order to obtain a state that lay on the SPEEDY model attractor. After this stabilization period, and starting from the last outcome from the model, we ran again the SPEEDY model for an additional 4-month period (hereafter, CNTRL). This last simulation was used to represent the true state of the atmosphere (Figure 1). Error prone radiosonde observations were generated from this state by adding uncorrelated Gaussian random numbers to the true state every 6 hours. The observation error standard deviations used were 1.0 K, 1.0 m s^{-1}, 0.1 g kg^{-1} and 1 hPa for temperature, zonal and meridional winds, specific humidity and surface pressure, respectively. At the observing stations, temperature and winds are observed at all seven layers, and specific humidity is observed between layers 1-4. Because of very thin humidity, specific humidity was not assimilated at higher levels. The horizontal locations of the observations are shown in Figure 2.

In order to be able to use the LETKF, first we need to generate our initial ensemble of perturbations. In this study, we have generated such an ensemble by sampling from the time sequence of true states every 12 hours. In other words, ensemble member 1 is chosen as the true state at the end of the 1-year spin-up period (Figure 1). Then, ensemble member 2 is the true state corresponding to the true state 12 hours later than member 1, and member 3 is the true state 12 hours later than member 2, and so forth. It is important to note that although the ensemble member 1 corresponds to the true state at the time we are initializing the ensemble, this is no longer valid after the first DA cycle. Using this initial ensemble, we used the LETKF to assimilate globally distributed radiosonde observations (Kotsuki, et al., 2020). However, in order to assimilate these observations using the LETKF, it is important to bear in mind that the implementation of ensemble Kalman filters in realistic frameworks (i.e., use of relative small ensemble members $\sim\mathcal{O}(50)$) are characterized with two problems: forecast error covariance matrices (i) are rank deficient and (ii) contain spurious long-distance correlations. To ameliorate these deficiencies, covariance localisation techniques are employed. Hybrid variational schemes typically use model space **B**-localisation functions while LETKFs use observation space or **R**-localisation functions (Greybush et al., 2011). In the case of **B**-localisation, error correlations between distant model variables are attenuated by applying a separation distance dependent localisation function C_{ij}^{B} to the elements of the forecast error covariance matrix. In **R**-localisation, a separation distance dependent localisation function C_{ij}^{R} is applied to the elements of the inverse observation error covariance matrix $\mathbf{R}^{-1}$. Following Greybush et al. (2011), the LETKF used here employs a Gaussian-like horizontal localisation function C_{ij}^{R} given by

$$C_{ij}^{R} = \exp\left[-\frac{d_{ij}^{2}}{2L^{2}}\right] \qquad (11)$$

where d_{ij} is again the separation distance defined previously in the above sections and *L* is the localisation length scale. This localisation function is given compact support, setting it to zero when the separation distance exceeds a certain finite distance, also known as cutoff radius, which is defined as $r_{cutoff} = 2\sqrt{\frac{10}{3}}\,L = 3.65\,L$ and the observations outside of this radius are not assimilated. In the present study, a localisation length scale of 800 km was employed, meaning that observations beyond 2920 km of distance were not assimilated. A vertical localisation function is also employed following the same form used for the horizontal localisation function (Equation 11), using a vertical localisation length scale fixed at 0.1 log(Pa). In addition, in order to ensure that ensemble variance roughly matched forecast error variance, the adaptive multiplicative inflation method of Miyoshi (2011) was applied before each DA cycle.

This global heterogeneous network of radiosonde observations (Figure 2) was assimilated by the LETKF-SPEEDY with 20 ensemble members (20M-LETKF) using 6-hour DA cycles over a 2-month period (Figure 1). Figure 3 (solid curves) shows that the 20M-LETKF DA scheme markedly reduces Root Mean Square Error (RMSE) of the forecast in the first few cycles but then, after about a month (120 DA cycles), the RMSE of the system becomes stable. Furthermore, around this time, the RMSE becomes approximately equal to the ensemble standard deviation. Figure 3 clearly shows that after 2 months both the RMSE and spread-skill relationship have been stabilized.

Having "spun-up" the DA scheme in this first 2-month period, a subsequent 2-month LETKF-SPEEDY DA period was used to estimate aspects of the function $P_{ij}^{f}\left(P_{ij}^{ens}, d_{ij}\right)$ and its relationship to $B_{ij}^{c}\left(d_{ij}\right)$. To ensure that our sample of 2 months was large enough to accurately

approximate the true forecast error statistics, we performed two additional experiments where this 2-month period dataset was split into its 1st and 2nd month parts. Then, we computed our statistics for the first and second month separately and the results did not show significant differences between both months (see supplementary Figure S1 and Figure S2, respectively). These results are consistent with the hypothesis that our data set is large enough to ensure stable statistics. It is noteworthy that ideally, if one where to employ our method to create a Hybrid covariance model for operational use, one would use statistics that were as relevant to the time of year and region as possible. In our experiment, we did not attempt to account for the fact that the weights in the summer (spring) hemisphere are probably different to those in the winter (autumn) hemisphere; i.e., to explore seasonality we would need to allow the weights to be a function of region. However, since such seasonality issues are peripheral to the main foci of this paper, we do not pursue them here.

To enable a test of the sensitivity of results to ensemble size, we repeated the above procedure using an 80-ensemble member LETKF (80M-LETKF) instead of the 20M-LETKF. As expected, using the 80M-LETKF led to a more rapid reduction of errors and also a lower stabilized value of RMSE (Figure 3, dashed curves).

4.2 Computations to determine $P_{\bar{y}}^{f}\left(P_{\bar{y}}^{ens}, d_{\bar{y}}\right)$ for temperature variables

For the sake of brevity, in this paper we limit our consideration to that of error covariances of forecasts of the temperature field at model vertical level 4, lying between 57.5 S and 57.5 N and corresponding to the vertical pressure level of 500 hPa. Thus, in our case, the forecast error vector ε^{n} introduced in Section 2 only pertains to the list of variables defining the temperature field at level 4. Along lines of constant longitude, the South-North direction SPEEDY grid spacing is constant with a grid point separation of *~412* km. This fact enables us to further

simplify our analysis by only considering forecast error covariances of temperature variables having the same longitude. For this subset of covariances, the distances $d_{ij}(n) = n \cdot 412\ km \quad \{n \in \square \quad | 0 \le n \le 8\}$ between variables are always some multiple of the number of grid points separating the variables along lines of constant latitude. In addition, we only consider covariances between grid points whose separation distance are less than 3500 km. These restrictions meant that, for the n^{th} forecast, we only needed to save the data triplets $\left[e_{ij}^n, \left(P_{ij}^{ens}\right)^n, d_{ij}^n \right]$ between 57.5 S and 57.5 N pertaining to products e_{ij}^n of temperature forecast errors for which both temperature variables lie on model level 4 and for which both temperature variables lie on the same latitude line and for which the separation distance is less than 3500 km. All of the triplets $\left[e_{ij}^n, \left(P_{ij}^{ens}\right)^n, d_{ij}^n \right]$ satisfying these constraints were collected for all of the N=242 DA cycles of the 2-month test period. These triplets were then grouped into 9 sets, corresponding to variable separation distances of 0 km, 412 km, 824 km, 1236 km, 1648 km, 2060 km, 2472 km, 2884 km and 3296 km. $B_{ij}^c\left[\left(d_{ij}\right)_{\text{target}}\right]$ for each of these separation distances was then obtained by taking the average of the error products e_{ij} in each of these 9 sets.

To obtain $\overline{P_{ij}^f\left[\left(P_{ij}^{ens}\right)_{\text{target}}, \left(d_{ij}\right)_{\text{target}}\right]}$ each of these 9 sets of triplets were then split into subsets based on the ensemble-based prediction $\left(P_{ij}^{ens}\right)$ of the flow dependent covariance between the forecast errors of the i^{th} and j^{th} model variable. Specifically, this was done by taking one of the subsets of triplets of $\left[e_{ij}, \left(P_{ij}^{ens}\right), d_{ij} \right]$ corresponding to a fixed separation distance and then ordering this subset of triplets from the triplet having the smallest value of $\left(P_{ij}^{ens}\right)$ to the triplet having the largest value of $\left(P_{ij}^{ens}\right)$. This ordered list of triplets having the

same $\boldsymbol{d}_{\bar{y}}$ values was then split into 20 approximately equally populated bins each having similar values of $\left(P_{\bar{y}}^{ens}\right)$. The bin-average of the error products $\boldsymbol{e}_{\bar{y}}$ within this bin is then used to obtain our empirical estimate $\overline{P_{ij}^{f}\left[\left(P_{ij}^{ens}\right)_{\text{target}},\left(d_{ij}\right)_{\text{target}}\right]}$ of $\boldsymbol{P}_{\bar{y}}^{\boldsymbol{f}}\left[\left(\boldsymbol{P}_{\bar{y}}^{\boldsymbol{ens}}\right)_{\textbf{target}},\left(\boldsymbol{d}_{\bar{y}}\right)_{\textbf{target}}\right]$.

Note that in the limit of an infinite number of DA cycles, one could make the ranges of $\left(P_{\bar{y}}^{ens}\right)$ and $\boldsymbol{d}_{\bar{y}}$ within each bin infinitesimally small while still maintaining a subset of data triplets within each bin. In this limit, $\boldsymbol{P}_{\bar{y}}^{\boldsymbol{f}}\left[\left(\boldsymbol{P}_{\bar{y}}^{\boldsymbol{ens}}\right),\left(\boldsymbol{d}_{\bar{y}}\right)\right]=\overline{P_{ij}^{f}\left[\left(P_{ij}^{ens}\right)_{\text{target}},\left(d_{ij}\right)_{\text{target}}\right]}$.

5. Results

5.1 Empirical results and interpretation

Each panel of Figure 4 pertains to a different $\left(\boldsymbol{d}_{\bar{y}}\right)_{\textbf{target}}$ value. The blue dots on each panel give $\overline{P_{ij}^{f}\left[\left(P_{ij}^{ens}\right)_{\text{target}},\left(d_{ij}\right)_{\text{target}}\right]}$ and thus can be used to estimate the continuous function $\overline{P_{ij}^{f}\left[\left(P_{ij}^{ens}\right),\left(d_{ij}\right)_{\text{target}}\right]}$ for a fixed value of $\left(\boldsymbol{d}_{\bar{y}}\right)_{\textbf{target}}$. These panels show that, in general, $\overline{P_{ij}^{f}\left[\left(P_{ij}^{ens}\right),\left(d_{ij}\right)_{\text{target}}\right]}$ is an approximately linear function of $\left(\boldsymbol{P}_{\bar{y}}^{\boldsymbol{ens}}\right)$. Hence, for simplicity, linear regression was applied to obtain a best guess linear function of the form

$$\overline{P_{ij}^{f}\left[\left(P_{ij}^{ens}\right),\left(d_{ij}\right)_{\text{target}}\right]}=a_{ij}\left(P_{ij}^{ens}\right)+b_{ij} \qquad (12)$$

where the terms a_{ij} and b_{ij} refers to the slope and to the intersection of the adjusted regression line with the ordinate axis, respectively. It is important to note that these regression coefficients depend on the variable separation distance d_{ij} between the i^{th} and j^{th} grid points and hence are different for each panel.

To better see the relationship between Equation 12 and Hybrid error covariance models, note that, without loss of generality, we can let $b_{ij} = g_{ij} B_{ij}^{c}\left[\left(d_{ij}\right)_{\text{target}}\right]$ provided that $g_{ij} = \dfrac{b_{ij}}{B_{ij}^{c}\left[\left(d_{ij}\right)_{\text{target}}\right]}$. Taking this into account, the regression equation becomes:

$$\overline{P_{ij}^{f}\left[\left(P_{ij}^{ens}\right),\left(d_{ij}\right)_{\text{target}}\right]} = a_{ij}\left(P_{ij}^{ens}\right) + g_{ij} B_{ij}^{c}\left[\left(d_{ij}\right)_{\text{target}}\right] \qquad (13)$$

This equation is like a Hybrid error covariance model in that it is a weighted average of P_{ij}^{ens} and $B_{ij}^{c}\left[\left(d_{ij}\right)_{\text{target}}\right]$ but because g_{ij} is a function of the separation distance it is also fundamentally different.

If ensemble perturbations were a sample from the distribution of true forecast errors, the average of all the ensemble covariances $\left\langle\left(P_{ij}^{ens}\right)\right\rangle$ in all of the bins having the same $\left(d_{ij}\right)_{\text{target}}$ value over the 2-month period would be approximately equal to $B_{ij}^{c}\left[\left(d_{ij}\right)_{\text{target}}\right]$. If this does not occur, it is a simple matter to correct the problem prior to DA by creating an adjusted ensemble covariance given by $\left(P_{ij}^{ens}\right)^{adj} = \dfrac{\overline{B_{ij}^{c}\left(d_{ij}\right)}}{\left\langle\left(P_{ij}^{ens}\right)\right\rangle}\left(P_{ij}^{ens}\right)$. One can then readily compute the

points $\overline{P_{ij}^{f}\left[\left(P_{ij}^{ens}\right)_{\text{target}}^{adj},\left(d_{ij}\right)_{\text{target}}\right]}$ (shown as red dots in Figure 4). Furthermore, using this adjusted covariance, Equation 13 can be rewritten as:

$$\overline{P_{ij}^{f}\left[\left(P_{ij}^{ens}\right),\left(d_{ij}\right)_{\text{target}}\right]}=h_{ij}\left(P_{ij}^{ens}\right)^{adj}+g_{ij}B_{ij}^{c}\left[\left(d_{ij}\right)_{\text{target}}\right] \tag{14}$$

where setting $\frac{h_{ij}}{a_{ij}}=\frac{\left\langle\left(P_{ij}^{ens}\right)\right\rangle}{B_{ij}^{c}\left(d_{ij}\right)}$ ensures that the left-hand side of Equation 14 is identical to the left-hand side of Equation 13. The simplifying aspect of using Equation 14, rather than Equation 13, can be seen by taking the average over the entire test period of both sides of Equation 14. This yields

$$\begin{aligned}B_{ij}^{c}\left[\left(d_{ij}\right)_{\text{target}}\right]&=\left\langle\overline{P_{ij}^{f}\left[\left(P_{ij}^{ens}\right)^{adj},\left(d_{ij}\right)_{\text{target}}\right]}\right\rangle=h_{ij}\left\langle\left(P_{ij}^{ens}\right)^{adj}\right\rangle+g_{ij}B_{ij}^{c}\left[\left(d_{ij}\right)_{\text{target}}\right]\\&=h_{ij}B_{ij}^{c}\left[\left(d_{ij}\right)_{\text{target}}\right]+g_{ij}B_{ij}^{c}\left[\left(d_{ij}\right)_{\text{target}}\right]\\&=\left(h_{ij}+g_{ij}\right)B_{ij}^{c}\left[\left(d_{ij}\right)_{\text{target}}\right]\end{aligned} \tag{15}$$

The only way Equation 15 can be satisfied is if $h_{ij}+g_{ij}=1$. Thus, adjusting the ensemble covariances in this way leads to an equation in which the coefficients are guaranteed to sum to unity.

Each panel of Figure 4 gives the best-fit regression equations (Equation 13 and 14) along with the associated values of a_{ij}, g_{ij} and h_{ij} for a particular $\left\langle d_{ij}\right\rangle_{\text{target}}$ value. Comparison of the a_{ij} and h_{ij} values in the differing panels shows how as the distance $\left\langle d_{ij}\right\rangle_{\text{target}}$ between i^{th}

and j^{th} grid points increases, the coefficients a_{ij} and h_{ij} decrease. This decrease may be viewed as an empirically estimated ensemble covariance *localisation* function. In contrast, the coefficient g_{ij} of the static covariance $B_{ij}^{c}\left[\left(d_{ij}\right)_{\text{target}}\right]$ increases as the separation distance increases. This increase may be viewed as an empirically determined *delocalisation* function.

To see this quantitatively, the coefficients of these terms for the case of the variances (i.e., $\left\langle d_{ij}\right\rangle_{bin}=0$) shows values of $a_{ij}=0.91$ and $g_{ij}=0.16$ for the $\left(P_{ij}^{ens}\right)$ and $B_{ij}^{c}\left[\left(d_{ij}\right)_{\text{target}}\right]$, respectively (Figure 4a), showing that the ensemble variance contributes to our empirical estimation of the actual forecast error variance more than the climatological error variance. However, as the separation distance is increased, the weight on the climatological covariance increases while that on the ensemble covariance decreases. For example at 1648 km, these weights are given by $a_{ij}=0.23$ for $\left(P_{ij}^{ens}\right)$ and $g_{ij}=0.84$ for $B_{ij}^{c}\left[\left(d_{ij}\right)_{\text{target}}\right]$ (Figure 4d). This increase of the weight on $B_{ij}^{c}\left[\left(d_{ij}\right)_{\text{target}}\right]$ with separation distance is *not* a feature of current Hybrid covariance models.

Extending these results for the entire set of distances $d_{ij}(n)=n\cdot 412\ km\ \left\{n\in \square\ \middle|\ 0\le n\le 8\right\}$ allows us to investigate the shape and values associated with the full ensemble-based (static) covariance localisation (delocalisation) functions estimated empirically. In order to depict such localisation functions we have represented the coefficients a_{ij}, h_{ij} and g_{ij} associated with the respective $\left(P_{ij}^{ens}\right),\left(P_{ij}^{ens}\right)^{adj}$ and B_{ij}^{c} terms as a function of the distance (Figure 5). The mauve and brown lines on Figure 5a

respectively depict a_{ij} and h_{ij} as a function of separation distance for the 20-member ensemble. The shapes of these curves correspond to what may be viewed as empirically determined localisation functions for the ensemble covariance matrix. The green line on Figure 5a depicts g_{ij} as a function of separation distance for the 20-member ensemble. It may be viewed as an empirically determined *delocalisation* function for the climatological covariances B_{ij}^{c}. From this result we can also identify at which separation distance the climatological term becomes more important than the ensemble term. In the case of using a 20-member ensemble system, such separation distance is approximately ~1500 km. Figure 5b is identical to Figure 5a except it pertains to the 80-member ensemble. As would be anticipated, the empirical localisation and delocalisation functions are much broader scale for this larger ensemble size. The distance at which the climatological and ensemble weights become equal is now ~2200 km. This distance is significantly larger than the corresponding 1500 km separation distance found using the 20-member ensemble-based covariance matrix.

In order to provide some more insight about the performance of our system, we have also computed the ratio between the $\left(P_{ij}^{ens}\right)$ and $\left(P_{ij}^{ens}\right)^{adj}$ localisation functions $\frac{h_{ij}}{a_{ij}} = \frac{\left\langle \left(P_{ij}^{ens}\right)\right\rangle}{B_{ij}^{c}\left(d_{ij}\right)}$ for the 20- and 80-member ensembles, respectively (Figure 5c). In other words, Figure 5c gives the ratio $\frac{\left\langle \left(P_{ij}^{ens}\right)\right\rangle}{B_{ij}^{c}\left(d_{ij}\right)}$ as a function of separation distance d_{ij} for both the 20- and 80-member ensemble cases (Figure 5c). Values of $\frac{\left\langle \left(P_{ij}^{ens}\right)\right\rangle}{B_{ij}^{c}\left(d_{ij}\right)} > 1$ indicate that the ensemble is over-estimating the covariance on average whereas $\frac{\left\langle \left(P_{ij}^{ens}\right)\right\rangle}{B_{ij}^{c}\left(d_{ij}\right)} < 1$ indicates that the ensemble is

underestimating the covariance on average. Figure 5c shows that at relatively short distances the 80-member ensemble slightly overestimates the actual covariance while with 20 members it slightly underestimates the actual forecast error covariance, on average. Beyond a separation distance of 1000 km, the value of $\frac{\left\langle \left(P_{ij}^{ens} \right) \right\rangle}{B_{ij}^{c}\left(d_{ij} \right)}$ exhibits deviations from unity that are greater than 20%. Presumably, this is because the climatological covariance in the denominator becomes close to zero at such distances thus making the ratio $\frac{\left\langle \left(P_{ij}^{ens} \right) \right\rangle}{B_{ij}^{c}\left(d_{ij} \right)}$ highly sensitive to limitations of the sample size.

5.2 Discussion of the unexpected behaviour of $P_{\bar{ij}}^{f}\left[\left(P_{\bar{ij}}^{ens} \right), \left(d_{ij} \right)_{\text{target}} \right]$ for $P_{\bar{ij}}^{ens}$ negative but near zero.

A striking result seen in Figure 4c and Figure 4d is that the blue dots given by $\overline{P_{ij}^{f}\left[\left(P_{ij}^{ens} \right)_{\text{target}}, \left(d_{ij} \right)_{\text{target}} \right]}$ imply a negatively sloped $P_{\bar{ij}}^{f}\left[\left(P_{\bar{ij}}^{ens} \right), \left(d_{ij} \right)_{\text{target}} \right]$ curve for $P_{\bar{ij}}^{ens}$ negative but near zero. In this neighbourhood, the forecast error covariance becomes more positive as the ensemble covariance becomes more negative – the opposite of what is assumed by all ensemble-based DA schemes including those that use Hybrid covariances!

What could be causing this unanticipated behaviour? At first, we suspected an error in our binning code when we saw this result. However, after further reflection we hypothesize that this feature is related to the use of spurious ensemble sample-covariances and variances in the LETKF DA scheme. An immediate prediction of our hypothesis is that the effect would be

diminished in systems where the P_{ij}^{ens} values were more accurate than the 20-member ensemble LETKF used to produce Figure 4. To assess this prediction, we repeated our experiments in exactly the same way but used an 80-member ensemble instead of a 20-member ensemble in the expectation that the 80-member ensemble would predict P_{ij}^{f} more accurately than the 20-member ensemble. Figure 6 is identical to Figure 4 except that it is based on data generated by an 80-member LETKF DA cycle rather than a 20-member LETKF experiment. (Note that neither the vertical nor horizontal localisation length scale were changed for this experiment. The only element changed was the size of the ensemble). Figure 6c and Figure 6d show that the negative slope behaviour for negative but near zero P_{ij}^{ens} is much smaller for the 80-member ensemble than the 20-member ensemble. This result is consistent, at least in part, with our initial hypothesis that the non-monotonicity behaviour observed at zero is consequence of spurious ensemble covariances. Admittedly, from the simple comparison of Figure 6 and Figure 4, it is not clear whether it is really the poorer quality of the 20-member LETKF analysis or whether this behaviour comes simply from the large sampling error computing the ensemble covariance. In order to shed light on this point, we redid Figure 6 but this time using a 20-member sub-sample of the 80-member ensemble. Results from this new figure (not shown) are very similar to the ones obtained in Figure 4. This shows that the spurious minimum observed at zero is not because of a fundamental sub-optimality in the gain matrix used in the cycling 20M-LETKF, but is simply due to spurious sample correlations associated with having a reduced ensemble (i.e., 20 members).

Small ensemble sizes cause spurious fluctuations in both the ensemble variances and the ensemble correlations. Furthermore, one would expect the fluctuations of ensemble variances and ensemble correlations to have a strong degree of independence when the underlying true error correlation is near zero – as it is in Figures 4c and 4d. To indicate the importance of

spurious ensemble variance fluctuations on the shapes of the curves in Figures 4 and 6, we redid Figure 4 but this time we binned errors based on ensemble *correlations*, rather than ensemble covariances. The curves resulting from the correlation-based binning are shown in Figure 7. Unlike in Figure 4, the curves in Figure 7 have no minimum at the zero-ensemble correlation. (Figure 7 only shows results associated to panel c) and d) from Figure 4, which are the situations where the local minimum is observed).

The result indicates that the spurious fluctuation of ensemble variances in situations where the ensemble correlations are near zero is a significant contributor to the local minima seen in Figures 4c and 4d. It also encourages more research to find even better predictors of the actual forecast error covariance than the simple linear function of ensemble covariance given by the Hybrid formulation – particularly when the ensemble size has ~ 20 members rather than ~ 100 members.

5.3 Similarity of Hybrid and climatological covariance matrices to $\left\{\mathbf{P}^f\right\}_{ij} = \left\langle \varepsilon_i^f \varepsilon_j^f \,|\, \left\{\mathbf{P}_{ens}^f\right\}_{ij} \right\rangle$

The typical Hybrid covariance model used in operational variational schemes takes the form

$$\mathbf{P}^H = a_H \mathbf{C} \circ \mathbf{P}^{ens} + b_H \mathbf{B}^c \qquad (16)$$

where a_H and b_H are scalars that may be a function of geographic location and **C** is a localisation matrix. Usually this matrix is based on a prescribed localisation length scale. Note that the Hybrid covariance model given by Equation 16 has a key structural difference to the

linear empirical model given by Equation 13. In Equation 16, the coefficient b_H of the climatological covariance matrix is a scalar and does not vary with variable separation distance. In contrast, in Equation 13, the corresponding coefficient g_{ij} does vary with separation distance. This fundamental difference can also be seen by writing Equation 13 in the matrix form

$$\overline{\mathbf{P}^f\left[\left(\mathbf{P}^{ens}\right)\right]} = \mathbf{A} \odot \mathbf{P}^{ens} + \mathbf{G} \odot \mathbf{B}^c \qquad (17)$$

where **A** is the localisation matrix whose localisation function is based on the values of a_{ij} for differing separation distances while **G** is the *delocalisation* matrix whose delocalisation function is based on the values of g_{ij} for differing separation distances.

For purposes of pictorially highlighting the differences in the covariance models given by Equation 16 and Equation 17 we constructed our own approximation of the typical Hybrid covariance model given by Equation 16. It is common for operational centres to base the localisation matrix **C** used in Hybrid variational schemes on the localisation function that gives good performance in ensemble DA schemes. To mimic this approach, for our approximation to Equation 16 we chose C_{ij}^B to be based on the Gaussian-like localisation function described in Equation 11, with a 800 km localisation length scale, which is applied on the inverse observation error covariance matrix used in the LETKF.

Operational centres also pay a large amount of attention to predicting forecast error variance as accurately as possible. The line of best fit to the variance data in Figure 4a (blue labeled linear regression equation), shows that choosing $a_H = 0.91$ and $b_H = 0.16$ yields the

linear combination of ensemble variances and climatological variances that optimally predicts forecast error variance. For this reason, we choose those values for these parameters.

For two different grid-points at the initial time of our DA test period, Figure 8 compares commonly used forecast error covariance models against the empirically derived estimate of $\overline{P_{ij}^{f}\left[\left(P_{ij}^{ens}\right),\left(d_{ij}\right)_{\text{target}}\right]}$. Visual inspection of Figure 8 suggests that over these two cases, the Hybrid (green line) is significantly closer to $\overline{P_{ij}^{f}\left[\left(P_{ij}^{ens}\right),\left(d_{ij}\right)_{\text{target}}\right]}$ than either the localized ensemble covariance (dashed blue line) or the climatological covariance matrix (red line).

To quantitatively assess the benefits of each of these approaches we used the Root Mean Square Distance (RMSD) verification score over all considered distances $\left(d_{\bar{ij}}\right)_{\text{target}}$ defined by:

$$\overline{RMSD}^{\left(d_{\bar{ij}}\right)}=\sqrt{\frac{1}{N_d}\sum_{d_{\bar{ij}}}\left[P_{\bar{ij}}^{guess}\left\{\left(d_{\bar{ij}}\right)_{\text{target}}\right\}-\overline{P_{\bar{ij}}^{f}\left\{P_{\bar{ij}}^{ens},\left(d_{\bar{ij}}\right)_{\text{target}}\right\}}\right]^{2}} \qquad (18)$$

where N_d is the number of the 9 different considered distances $\left(d_{\bar{ij}}\right)_{\text{target}}$ (see Section 4.2 for further details) and $P_{\bar{ij}}^{guess}\left(d_{\bar{ij}}\right)$ denotes one of these frequently used forecast error covariance models (i.e., $B_{\bar{ij}}^{c}$, $C_{\bar{ij}}^{B}\left(P_{\bar{ij}}^{ens}\right)$, $P_{\bar{ij}}^{ens}$ and the traditional Hybrid given by Equation 16). In practice, to obtain the $\overline{RMSD}^{\left(d_{\bar{ij}}\right)}$ between the different forecast error covariance models $P_{\bar{ij}}^{guess}\left(d_{\bar{ij}}\right)$ and the traditional Hybrid, we proceed as follows: (i) we compute differences $\left(P_{ij}^{guess}\left(d_{ij}\right)_{\text{target}}-\overline{P_{ij}^{f}\left[\left(P_{ij}^{ens}\right),\left(d_{ij}\right)_{\text{target}}\right]}\right)$ for each of the 9 different distances corresponding to a

specific DA cycle (ii) we take the square of these 9 different values, (iii) we compute the average of these values, and finally (iv) we take the square root.

The $\overline{RMSD}^{(d_y)}$ values corresponding for each of the abovementioned $P_{ij}^{guess}\left(d_{ij}\right)$ terms are given in the legend of each figure. They show that the Hybrid form is associated with the minimum value of $\overline{RMSD}^{(d_y)}$, confirming the benefits of the Hybrid form in comparison with the other approaches results (Figure 8). It is noteworthy that the grid-points selected in Figure 8 correspond to two different situations. The ensemble variance and corresponding forecast error variance given the ensemble variance is much larger than the climatological variance for the grid point corresponding to Figure 8a. In this case, for the Hybrid form, $\overline{RMSD}^{(d_y)} = 0.044\ \mathrm{K}$, which is slightly lower (better) than the $\overline{RMSD}^{(d_y)} = 0.046\ \mathrm{K}$ corresponding to the localized ensemble-based covariance and much lower than the $\overline{RMSD}^{(d_y)} = 0.163\ \mathrm{K}$ corresponding to the climatological covariance. At the grid point pertaining to Figure 8b, the ensemble variance and corresponding forecast error variance given the ensemble variance is about the same as the climatological variance. In this case, the $\overline{RMSD}^{(d_{ij})}$ for the Hybrid, localized ensemble covariance and climatological covariance are given by 0.026 K, 0.037 K and 0.034 K, respectively. Thus, for both grid-points the Hybrid covariance model is significantly closer to the empirically determined forecast error covariance function than either the localized ensemble covariance model or the climatological error covariance model.

To quantify the extent to which the specific improvements seen in Figure 8 are representative of all cases, we averaged over many grid points and many DA times in the following way. Specifically, we use a slight variation of Equation 18:

$$\overline{RMSD}^{(t,\lambda,\phi)}\left(d_{ij}\right)=\sqrt{\frac{1}{N_t N_\lambda N_\phi}\sum_{t}\sum_{\lambda,\phi}\left[P_{ij}^{guess}\left\{\left(d_{ij}\right)_{\text{target}}\right\}-\overline{P_{ij}^{f}\left\{P_{ij}^{ens},\left(d_{ij}\right)_{\text{target}}\right\}}\right]^2} \tag{19}$$

where N_t, N_λ, N_ϕ are the number of time steps, latitude and longitude grid-points, respectively. Note that Equation 19, before taking the square root, it averages the square difference between the covariance models and the Hybrid term over all the grid-points and also over time, instead of averaging over the 9 different considered distances $\left(d_{ij}\right)_{\text{target}}$ as was considered in Equation 18. In practice, to compute $\overline{RMSD}^{(t,\lambda,\phi)}\left(d_{ij}\right)$ for each separation distance considered $\left(d_{ij}\right)_{\text{target}}$, (i) we take all the corresponding pair values $\left(P_{ij}^{guess}\left(d_{ij}\right)_{\text{target}}-\overline{P_{ij}^{f}\left[\left(P_{ij}^{ens}\right),\left(d_{ij}\right)_{\text{target}}\right]}\right)$ associated to all the grid-points with latitudes between 57.5 S and 57.5 N and separation distance $\left(d_{ij}\right)_{\text{target}}$, over the 2-month period of DA simulation (ii) we square such differences for each of the corresponding forecast error covariance models (iii) we temporally and spatially average these quantities, and finally (iv) we take the square root. Results of these computations as a function of $\left(d_{ij}\right)_{\text{target}}$, for each of the covariance models $P_{ij}^{guess}\left(d_{ij}\right)$, is showed in Figure 9.

On the one hand, Figure 9 shows how the $\overline{RMSD}^{(t,\lambda,\phi)}\left(d_{ij}\right)$ associated with the localized ensemble covariance function is smaller than the climatological covariance function for relatively small geographical distances (i.e., <1500 km), but as the distance increases, the

climatological covariance function provides better information reducing the $\overline{RMSD}^{(t,\lambda,\phi)}\left(d_{ij}\right)$ below the localized ensemble-based covariance. On the other hand, the traditional Hybrid formulation provides the smallest $\overline{RMSD}^{(t,\lambda,\phi)}\left(d_{ij}\right)$scores among all the other commonly used forecast error covariance models, at least considering distances smaller than 2200 km. However, if we consider grid point distances greater than 2200 km, the $\overline{RMSD}^{(t,\lambda,\phi)}\left(d_{ij}\right)$of the climatological approximation to the actual forecast error covariance function becomes smaller than the Hybrid approximation. This result clearly indicates that as the variable separation distance increases it is necessary to increase the weight on the climatological term, contrary to the traditional Hybrids which keep this weight constant.

In conclusion, our empirical findings have shown that the current Hybrid form used by many operational centres is a much better approximation to the actual covariance of forecast errors *given an ensemble covariance* than either the static climatological covariance or the localized ensemble-based covariance. This finding helps explain why operational centres have found such large forecast improvements when switching from a static error covariance model to a Hybrid forecast error covariance model. Another fascinating finding of our empirical study is that the *form* of current Hybrid error covariance models is fundamentally incorrect in that the weight given to the static covariance matrix is independent of the separation distance of model variables. Our results show that this weight should be an increasing function of this separation distance.

6. Summary and Discussion

Using a simplified Global Circulation Model and LETKF DA scheme, we have empirically derived and explored the relationship between specific ensemble temperature covariance values and the covariance of actual forecast errors corresponding to these specific values. To a first approximation, the covariance of actual forecast errors has been found to be a monotonically quasi-linear increasing function of the ensemble covariance. The slope of the line of best fit decreases as the distance between the variables increases. The assumption of a linear relationship is shown to be violated for small ensemble sizes and relatively large separation distances. In this regime, the covariance of actual forecast errors is *not* a monotonically increasing function of increasing ensemble covariance. Specifically, for ensemble covariances that are close to zero but negative, the covariance of actual forecast errors *decreases* as the corresponding ensemble covariance *increases*. As far as the authors are aware, this is the first study to point out this feature. We hypothesise that this departure from monotonicity is a consequence of spurious fluctuations of the ensemble sample-covariance and variances from sampling errors obtained using a small ensemble size (20 ensemble members). Consistent with this hypothesis, we find that quadrupling the size of the ensemble to 80 members recovers a monotonic relationship. The role of spurious ensemble variance fluctuations on the local minimum in these curves was highlighted by binning forecast errors based on ensemble correlations, rather than on ensemble covariances. Results showed that the minimum at zero disappears in this case. Our findings indicate that future research aimed at a deeper understanding of how spurious sample correlations and variances result in the minimums seen in Figure 4 might lead to even more accurate ensemble-based estimates of the true forecast error covariance given an ensemble covariance.

We have also shown that the traditional Hybrid form used by many operational centres is a much better approximation to the actual covariance of forecast errors given an imperfect ensemble covariance than either the static climatological covariance or the localized ensemble

covariance. This finding helps explain why operational centres have found such large forecast improvements when switching from a static error covariance model to a Hybrid forecast error covariance model.

Our empirically determined forecast error covariance model yields an empirically determined localisation function for ensemble covariances and a distance dependent weighting function (*a delocalization function*) for the climatological covariances. This empirical result is consistent with the Hybrid gain matrix formulation proposed by Flowerdew (2015). Our empirical findings support Flowerdew's (2015) argument that the current form of Hybrid error covariance models is fundamentally limited by the fact that the weight given to the static climatological error covariance matrix is independent of the separation distance of model variables. Our results show that this weight should be an increasing function of the separation distance. This behaviour causes the climatological covariance matrix to dominate the Hybrid covariance matrix at large separation distances. Although the present study has shed light on the improvement of Hybrid DA schemes, further research is needed to develop effective numerical strategies for implementing the more accurate form of the Hybrid revealed by this empirical study to variational and ensemble-based DA schemes. An immediate challenge is the fact that the moderation matrix associated with weights that increase with separation distance will have negative eigenvalues and hence an imaginary square root. Current techniques do not accommodate imaginary square roots. Nevertheless, there are a number of possible ways forward that future research will address.

ACKNOWLEDGEMENTS

D. S. Carrió and C.H. Bishop were supported by the ARC Centre of Excellence for Climate Extremes (CE170100023). This study was partly supported by the Japan Aerospace

Exploration Agency (JAXA) Precipitation Measuring Mission (PMM), the Japan Society for the Promotion of Science (JSPS) KAKENHI grant JP18H01549, and JST PRESTO MJPR1924. The authors also declare that they do not have any conflict of interest.

References

Anderson, J. L., 2001. An ensemble adjustment Kalman filter for data assimilation.. *Monthly weather review*, pp. 2884-2903.

Bishop, C. H., Etherton, B. J. & Majumdar, S. J., 2001. Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects.. *Monthly weather review*, pp. 420-436.

Bishop, C. H. & Satterfield, E., 2013. Hidden Error Variance Theory. Part I: Exposition and Analytic Model. *Monthly Weather Review*, Volume 141, pp. 1454-1468.

Bonavita, M., Torrisi, L. & Marcucci, F., 2008. The ensemble Kalman filter in an operational regional NWP system: preliminary results with real observations. *Q. J. R. Meteorol. Soc*, Volume 134, pp. 1733-1744.

Clayton, A., Lorenc, A. C. & Barker, D. M., 2013. Operational implementation of a hybrid ensemble/4D-Var global data assimilation system at the Met Office. *Q. J. R. Meteorol. Soc.*, Volume 139, pp. 1445-1461.

Courtier, P., Thépaut, J. N. & Hollingsworth, A., 1994. A strategy for operational implementation of 4D-Var, using an incremental approach. *Quarterly Journal of the Royal Meteorological Society*, pp. 1367-1387.

Daley, R., 1993. *Atmospheric data analysis.* s.l.:Cambridge University Press.

Evensen, G., 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics.. *J. Geophysic. Res.*, Volume 99.

Flowerdew, J., 2015. Towards a theory of optimal localisation. *Tellus A: Dynamic Meteorology and Oceanography*, 67(1), p. 25257.

Greybush, S. J. et al., 2011. Balance and ensemble Kalman filter localization techniques. *Monthly weather review*, Volume 139, pp. 511-522.

Hamill, T. M., 2006. Ensemble-based atmospheric data assimilation. *Predictability of weather and climate*, pp. 124-156.

Hamill, T. M. & Snyder, C., 2000. A Hybrid Ensemble Kalman Filter-3D Variational Analysis Scheme. *Mon. Weather Rev.*, Volume 128, pp. 2905-2919.

Hollingsworth, A. & Lönnberg, P., 1986. The statistical structure of short-range forecast errors as determined from radiosonde data Part II: The covariance height and wind errors. *Tellus A*, pp. 137-161.

Houtekamer, P. L. & Mitchell, H. L., 1998. Data assimilation using an ensemble Kalman filter technique. *Monthly Weather Review*, pp. 796-811.

Houtekamer, P. L. & Mitchell, H. L., 2001. A sequential ensemble Kalman filter for atmospheric data assimilation.. *Monthly Weather Review*, Volume 129, pp. 123-137.

Houtekamer, P. et al., 2005. Atmospheric data assimilation with an ensemble Kalman filter: Results with real observations. *Monthly Weather Review,* pp. 133:604-620.
Huang, B., Wang, X. & Bishop, C., 2019. The high-Rank Ensemble Transform Kalman Filter. *Monthly weather review,* Volume 147, pp. 3025-3043.
Hunt, B. R., Kostelich, E. J. & Szunyogh, I., 2007. Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Physica D: Nonlinear Phenomena,* Volume 230(1-2), pp. 112-126.
Kalnay, E., 2003. *Atmoshperic modeling, data assimilation and predictability.* s.l.:Cambridge university press.
Kotsuki, S., Pensoneault, A., Okazaki, A. & Miyoshi, T., 2020. Weight structure of the Local Ensemble Transform Kalman Filter: A case with an intermediate atmospheric general circulation model. *Quarterly Journal of the Royal Meteorological Society.*
Kuhl, D. D. et al., 2013. Comparison of Hybrid Ensemble/4DVar and 4DVar within within the NAVDAS-AR Data Assimilation Framework.. Mon. Weather Rev., Volume 141, pp. 2740-2758.
Ménétrier, B. & Auligné, T., 2015. Optimized Localization and Hybridization to Filter Ensemble-Based Covariances. *Monthly Weather Review,* 143(10), pp. 3934-3947.
Miyoshi, T., 2011. The Gaussian approach to adaptive covariance inflation and its implementation with the local ensemble transform Kalman filter. *Monthly Weather Review,* pp. 1519-1535.
Molteni, F., 2003. Atmospheric simulations using a GCM with simplified physical parametrizations. I: Model climatology and variability in multi-decadal experiments. *Climate Dynamics,* pp. 175-191.
Press, S. J., 2012. *Applied multivariate analysis: using Bayesian and frequentist methods of inference.*. 2 ed. s.l.:Robert E. Krieger Publishing Co.
Reich, S. & Cotter, C., 2015. *Probabilistic forecasting and Bayesian data assimilation.* s.l.:Cambridge University Press.
Snyder, C., Bengtsson, T., Bickel, P. & Anderson, J., 2008. Obstacles to high-dimensional particle filtering. *Monthly Weather Review,* pp. 4629-4640.
Van Leeuwen, P. J., 2010. Nonlinear data assimilation in geosciences: an extremely efficient particle filter. *Quarterly Journal of the Royal Meteorological Society,* pp. 1991-1999.
Van Leeuwen, P. J., 2012. Particle filter for the geosciences. *Advanced Data Assimilation for Geosciences: Lecture Notes of the Les Houches School of Physics: Special Issue.*
Van Leeuwen, P. J., 2019. Particle filters for high-dim ensional geoscience applications: A review.. *Quarterly Journal of the Royal Meteorological Society,* pp. 2335-2365.
Wang, X. & Lei, T., 2014. GSI_Based Four-Dimensional Ensemble-Variational (4DEnsVar) Data Assimilation: Formulation and Single-Resolution Experiments with Real Data for NCEP Global Forecast System. *Monthly Weahter Review,* Volume 142, pp. 3303-3325.
Whitaker, J. S. & Hamill, T. M., 2002. Ensemble data assimilation without perturbed observations. *Monthly weather review,* pp. 1913-1924.
Whitaker, J. S. & Hamill, T. M., 2002. Ensemble data assimilation without perturbed observations. *Monthly weather review,* Volume 130, pp. 1913-1924.
Wishart, J., 1928. The Generalised Product Moment Distribution in Samples from a Normal, Multivariate Population.. *Biometrika,* Volume 20, pp. 23-52.

APPENDIX

Proof that $\left\langle \varepsilon^f \varepsilon^{fT} \mid \mathbf{P}_{ens}^f \right\rangle = \left\langle \mathbf{P}^f(\mathbf{y}) \mid \mathbf{P}_{ens}^f \right\rangle$ using the Bayesian perpective

In a Bayesian framework, one considers an infinity of replicate systems having different states but all having the same climate. Each of the replicate systems have observing networks that take observations of the same variables at the same points in space-time using instruments that have the same distributions of random observation errors. The observation error vector incurred in each replicate system is random and independent of the observation errors that are incurred in the other replicate systems. In this framework, the true forecast error covariance matrix is the covariance of the errors ε^f in the ensemble mean forecast on each of the replicate systems that happened to have exactly the same observed values over the entire history of observations. (Note that this thought experiment accommodates the fact that the mean of practicable ensemble forecasts inevitably differs from the mean of the ensemble of true states having the same historically observed values; i.e. practical ensemble forecasts are imperfect representations of the actual distribution of true states given past observed values). Listing the past observed values in the vector **y**, we can define the true flow-dependent forecast error covariance matrix using

$$\mathbf{P}^f = \mathbf{P}^f(\mathbf{y}) = \left\langle \varepsilon^f \varepsilon^{fT} \mid \mathbf{y} \right\rangle = \int_{V_{\varepsilon^f}} \varepsilon^f \varepsilon^{fT} \rho\left(\varepsilon^f \mid \mathbf{y}\right) dV_{\varepsilon^f} \qquad \text{(A1)}$$

where $\rho\left(\varepsilon^f \mid \mathbf{y}\right)$ is the probability density of forecast errors ε^f of all the replicate systems that have the same $\mathbf{y}$ vector. Note that as the data assimilation cycle proceeds, more and more unique values of $\mathbf{P}^f(\mathbf{y})$ are generated. An infinity of such cycles would define a climatological pdf of $\mathbf{P}^f(\mathbf{y})$ values. Each specific value of $\mathbf{P}^f(\mathbf{y})$ may be viewed as a random draw from this climatological pdf of $\mathbf{P}^f(\mathbf{y})$. Along with each value of $\mathbf{P}^f(\mathbf{y})$, a random forecast error ε^f and a random $\mathbf{P}_{ens}^f$ also occur. Similarly, the cycling ensemble data assimilation scheme builds up a climatological pdf of ensemble covariances $\mathbf{P}_{ens}^f$. Under the Ergodic assumption, the climatalogical densities of true forecast error covariances $\rho_{\text{clim}}\left(\mathbf{P}^f(\mathbf{y}) \mid \mathbf{P}_{ens}^f\right)$ and forecast errors $\rho_{\text{clim}}\left(\varepsilon^f \mid \mathbf{P}_{ens}^f\right)$ *given* an approximately fixed value of $\mathbf{P}_{ens}^f$ are *identical* to the distribution of forecast errors given a *specific* instance of the same approximate value of $\mathbf{P}_{ens}^f$.

In such a system, the infinitesimal climatological probability of obtaining $\mathbf{P}^f(\mathbf{y})$ given $\mathbf{P}^f_{ens}$ is $\rho\left(\mathbf{P}^f(\mathbf{y})|\mathbf{P}^f_{ens}\right)\delta V_{\mathbf{P}^f(\mathbf{y})}$ where $\delta V_{\mathbf{P}^f(\mathbf{y})}$ is an infinitesimal part of the $\frac{1}{2}\left(n^2+n\right)$ dimensional volume defined by the $\frac{1}{2}\left(n^2+n\right)$ potentially unique elements of the symmetric nxn matrix $\mathbf{P}^f(\mathbf{y})$.

Given a specific $\mathbf{P}^f(\mathbf{y})$, the probability of obtaining the forecast error ε^f is given by

$\rho\left(\varepsilon^f|\mathbf{P}^f(\mathbf{y})\right)\delta V_{\varepsilon^f}$, where δV_{ε^f} is an infinitesimal part of the *n*-dimensional space corresponding to the *n*-elements of ε^f. Note that $\rho\left(\varepsilon^f|\mathbf{P}^f(\mathbf{y})\right)\delta V_{\varepsilon^f}$ is equivalent to $\rho\left(\varepsilon^f|\mathbf{y}\right)\delta V_{\varepsilon^f}$. Given a specific value of the object $\mathbf{P}^f_{ens}$, the climatological probability $\rho\left(\varepsilon^f,\mathbf{P}^f(\mathbf{y})|\mathbf{P}^f_{ens}\right)\delta V_{\varepsilon^f}\delta V_{\mathbf{P}^f(\mathbf{y})}$ of obtaining the forecast error ε^f and the true flow-dependent covariance $\mathbf{P}^f(\mathbf{y})$ is given by

$$\rho\left(\varepsilon^f,\mathbf{P}^f(\mathbf{y})|\mathbf{P}^f_{ens}\right)\delta V_{\varepsilon^f}\delta V_{\mathbf{P}^f(\mathbf{y})}=\left[\rho\left(\mathbf{P}^f(\mathbf{y})|\mathbf{P}^f_{ens}\right)\rho\left(\varepsilon^f|\mathbf{P}^f(\mathbf{y})\right)\right]\delta V_{\varepsilon^f}\delta V_{\mathbf{P}^f(\mathbf{y})} \qquad \text{(A2)}$$

To obtain the climatological probability $\rho\left(\varepsilon^f|\mathbf{P}^f_{ens}\right)\delta V_{\varepsilon^f}$ of ε^f given $\mathbf{P}^f_{ens}$ irrespective of the value of $\mathbf{P}^f(\mathbf{y})$, there are two possible approaches:

(i) Collect all ε^f values corresponding to the same $\mathbf{P}^f_{ens}$ over a very long time series (obviously, such a procedure ignores variations in the unknown $\mathbf{P}^f(\mathbf{y})$) and empirically define the probabilities $\rho\left(\varepsilon^f|\mathbf{P}^f_{ens}\right)\delta V_{\varepsilon^f}$), or

(ii) Integrate Equation (A2) over the climatological distribution of $\mathbf{P}^f(\mathbf{y})$ values to obtain

$$\begin{aligned}\rho\left(\varepsilon^f|\mathbf{P}^f_{ens}\right)dV_{\varepsilon^f}&=\int_{V_{\mathbf{P}^f(\mathbf{y})}}\rho\left(\varepsilon^f,\mathbf{P}^f(\mathbf{y})|\mathbf{P}^f_{ens}\right)\delta V_{\varepsilon^f}\delta V_{\mathbf{P}^f(\mathbf{y})}\\&=\int_{V_{\mathbf{P}^f(\mathbf{y})}}\left\{\left[\rho\left(\varepsilon^f|\mathbf{P}^f(\mathbf{y})\right)\rho\left(\mathbf{P}^f(\mathbf{y})|\mathbf{P}^f_{ens}\right)\right]\delta V_{\varepsilon^f}\right\}\delta V_{\mathbf{P}^f(\mathbf{y})} \qquad \text{(A3)}\end{aligned}$$

The first approach is the empirical sampling method used in this paper. The second approach is the analytic integral method. Provided an infinite sample size is used the values of $\rho\left(\varepsilon^f|\mathbf{P}^f_{ens}\right)\delta V_{\varepsilon^f}$ found from each method will be identical. Hence, the covariance of the two distributions must also be identical (along with all other moments). The empirical sampling

method of computing the covariance of the errors whose distribution is $\rho\left(\varepsilon^f \mid \mathbf{P}^f_{ens}\right)\delta V_{\varepsilon^f}$ is simply given by

$$\left\langle \varepsilon^f \varepsilon^{fT} \mid \mathbf{P}^f_{ens} \right\rangle = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \varepsilon^f_i \varepsilon^{fT}_i \qquad \text{(A4)}$$

where ε^f_i is the i^{th} realization of forecast error and N is the total number of forecast error samples. This paper uses the empirical sampling method (A4) to approximate $\left\langle \varepsilon^f \varepsilon^{fT} \mid \mathbf{P}^f_{ens} \right\rangle$. The analytic integral method of computing this exact same quantity is to use (A3) and the analytic definition of the covariance:

$$\begin{aligned}
\underbrace{\left\langle \varepsilon^f \varepsilon^{fT} \mid \mathbf{P}^f_{ens} \right\rangle}_{\sigma^2_f} &= \int_{V_{\varepsilon^f}} \varepsilon^f \varepsilon^{fT} \rho\left(\varepsilon^f \mid \mathbf{P}^f_{ens}\right) \delta V_{\varepsilon^f} \\
&= \int_{V_{\varepsilon^f}} \int_{V_{\mathbf{P}^f(\mathbf{y})}} \left\{ \left[\varepsilon^f \varepsilon^{fT} \rho\left(\varepsilon^f \mid \mathbf{P}^f\left(\mathbf{y}\right)\right) \rho\left(\mathbf{P}^f\left(\mathbf{y}\right) \mid \mathbf{P}^f_{ens}\right) \right] \delta V_{\varepsilon^f} \right\} \delta V_{\mathbf{P}^f(\mathbf{y})} \qquad \text{(A5)} \\
&= \int_{V_{\mathbf{P}^f(\mathbf{y})}} \int_{V_{\varepsilon^f}} \left[\varepsilon^f \varepsilon^{fT} \rho\left(\varepsilon^f \mid \mathbf{P}^f\left(\mathbf{y}\right)\right) \delta V_{\varepsilon^f} \right] \left\{ \left[\rho\left(\mathbf{P}^f\left(\mathbf{y}\right) \mid \mathbf{P}^f_{ens}\right) \right] \right\} \delta V_{\mathbf{P}^f(\mathbf{y})}
\end{aligned}$$

To see that $\left\langle \varepsilon^f \varepsilon^{fT} \mid \mathbf{P}^f_{ens} \right\rangle = \left\langle \mathbf{P}^f\left(\mathbf{y}\right) \mid \mathbf{P}^f_{ens} \right\rangle$ from Equation (A5), simply note that because $\mathbf{P}^f\left(\mathbf{y}\right) = \int_{V_{\varepsilon^f}} \left[\varepsilon^f \varepsilon^{fT} \rho\left(\varepsilon^f \mid \mathbf{P}^f\left(\mathbf{y}\right)\right) \delta V_{\varepsilon^f} \right]$, Equation (A5) simplifies to:

$$\begin{aligned}
\left\langle \varepsilon^f \varepsilon^{fT} \mid \mathbf{P}^f_{ens} \right\rangle &= \int_{V_{\mathbf{P}^f(\mathbf{y})}} \mathbf{P}^f\left(\mathbf{y}\right) \left\{ \left[\rho\left(\mathbf{P}^f\left(\mathbf{y}\right) \mid \mathbf{P}^f_{ens}\right) \right] \right\} \delta V_{\mathbf{P}^f(\mathbf{y})} \qquad \text{(A6)} \\
&= \left\langle \mathbf{P}^f\left(\mathbf{y}\right) \mid \mathbf{P}^f_{ens} \right\rangle \text{ as was required.}
\end{aligned}$$

FIGURE CAPTIONS:

Figure 1. Experimental configuration scheme used with SPEEDY and LETKF DA system. Light grey shaded areas represent the spin-up period from the climatic and DA simulations, respectively. Solid arrows indicate the evolution of the deterministic SPEEDY model. Dotted arrows represent the evolution of each ensemble member along the DA window, in which observations were assimilated every 6 hours.

Figure 2. Global spatial distribution of the radiosonde observations assimilated every DA cycle (6-hours). Pseudo-observation values are obtained from the truth simulation (CNTRL). In the same way, observational error values assigned at each location are obtained by adding random perturbations to the truth state.

Figure 3. Evolution of the spread-skill relationship obtained from the 20M-LETKF (blue lines) and 80M-LETKF (red lines) experiments, respectively, during a total period of 4 months. The first two months were intended to account for the spin-up effect related to the DA performance. Results showed in this study are obtained from the following two months (grey shaded area). RMSE and ensemble standard deviation are computed using a latitudinal-dependent weight function to account for the high-density of model grid points distributed near the poles. Grey shaded area indicates the 2-months period using the LETKF after DA spin-up.

Figure 4. Representation of the actual forecast error covariance as a function of a given (i) 20-member raw ensemble-based covariance matrix (blue dots) and (ii) an adjusted ensemble-based covariance matrix (red dots). Information in the legends show the linear regression coefficients between actual forecast error covariance and ensemble-based covariance obtained for the temperature field at ~500 hPa using the standard form (Equation 13) and the adjusted version (Equation 14). Climatological means associated with , $\left(P_{\bar{i}\bar{j}}^{ens}\right)$, $\left(P_{\bar{i}\bar{j}}^{ens}\right)^{adj}$ and $B_{\bar{i}\bar{j}}^{c}$ are depicted at the top-left corner of each panel.

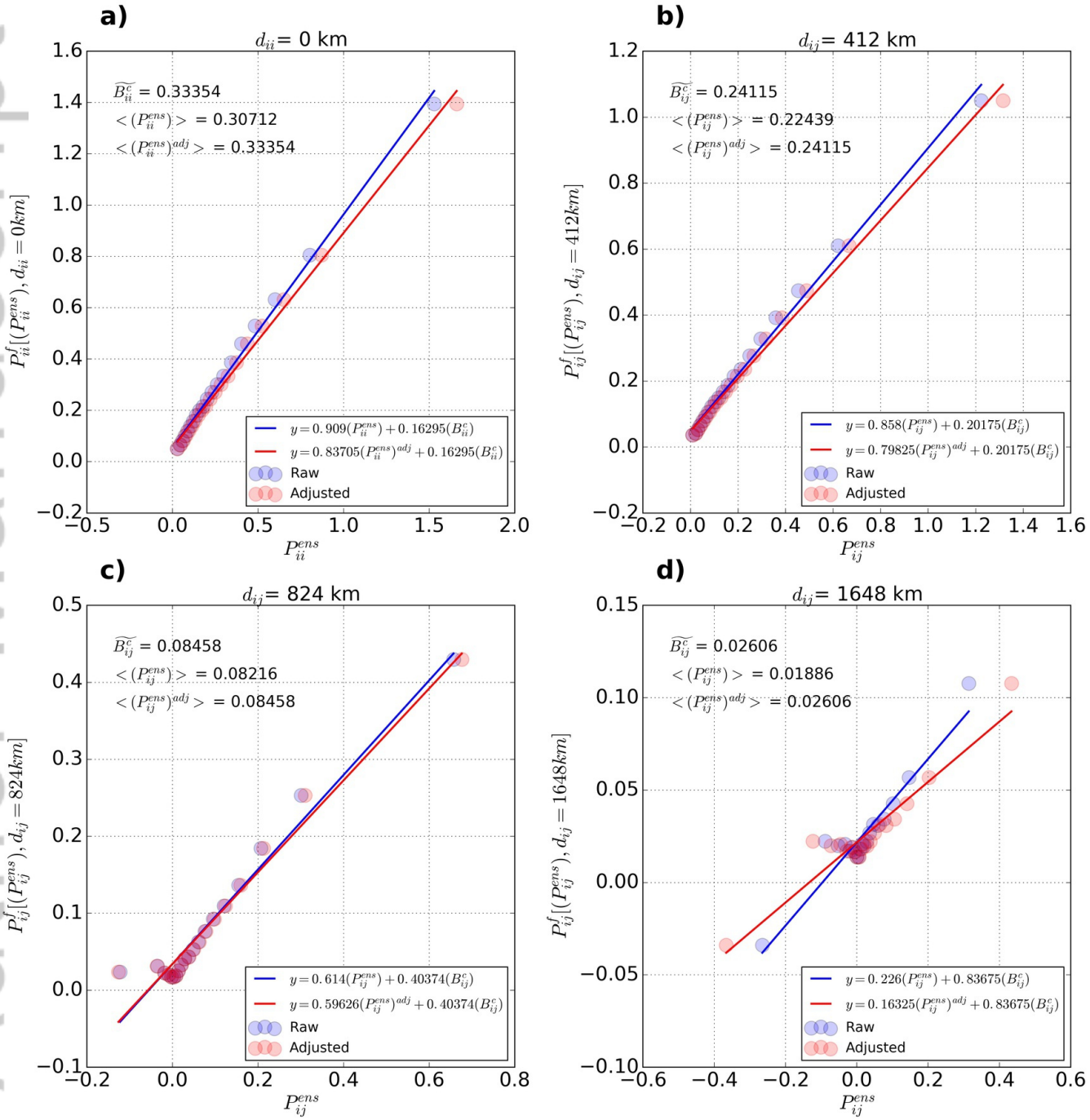
Figure 5. Representation of the empirical ensemble-based localisation function, associated to both $\left(P_{\bar{i}\bar{j}}^{ens}\right)$ and $\left(P_{\bar{i}\bar{j}}^{ens}\right)^{adj}$, together with the climatological deallocation function, associated to $B_{\bar{i}\bar{j}}^{c}$, as a function of the separation distance for the a) 20-member and b) 80-member ensemble. *Panel c) shows the ratio $h_{\bar{i}\bar{j}} / a_{\bar{i}\bar{j}}$ in function of the separation distance for the 20-member (solid line) and 80-member (dashed line) ensembles.*

Figure 6. Same as Figure 4 but for the case of the 80M-LETKF.

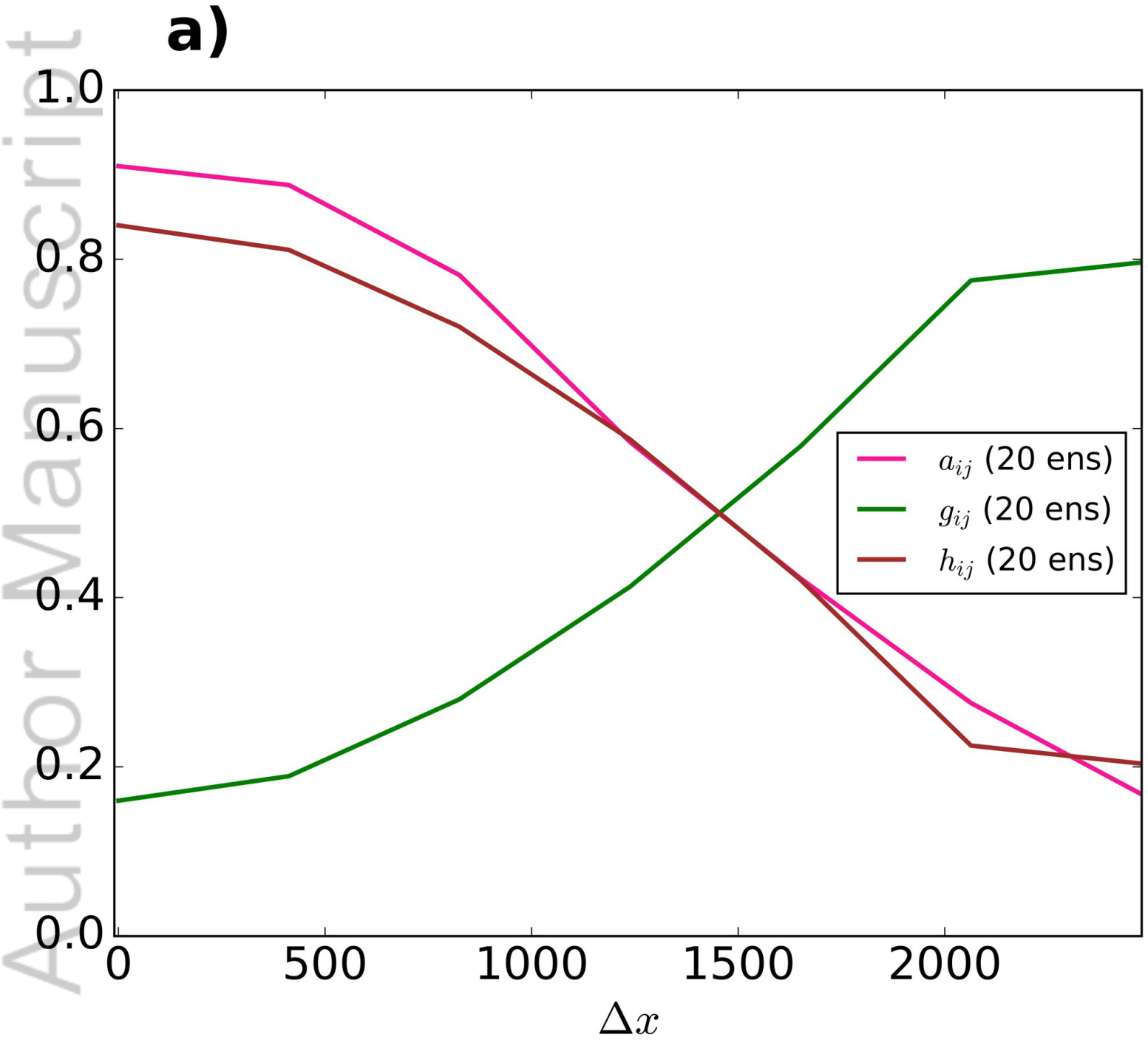
Figure 7. Analogous to Figure 4. Here, the forecast error covariance is binned according to their ensemble sample-correlation rather than their ensemble sample-covariance. Climatological means associated with , $\left(C_{ij}^{ens}\right)$ *and* B_{ij}^{c} *are also depicted at the top-left corner of each panel.*

Figure 8. Representation of the terms: B_{ij}^{c} , $\left(P_{ij}^{ens}\right)$, $C_{ij}^{B}\left(P_{ij}^{ens}\right)$ *and the traditional Hybrid form* $\overline{P_{ij}^{f}\left[\left(P_{ij}^{ens}\right),\left(d_{ij}\right)_{\text{target}}\right]}$ *in function of the physical distance* d_{ij} *(in km) for the model grid-point a) i=12, j=24 and b) i=28, j=52 at the initial time of the simulation for an ensemble of 20 member. Note that the overtilde in the true error covariance term is missing in the legend because of aspect reasons. In addition, it is also depicted the* $\overline{RMSD}^{\left(d_{ij}\right)}$ *which is the square root of the square differences of each of the covariance models with the Hybrid, averaged over all the distances* d_{ij} considered .

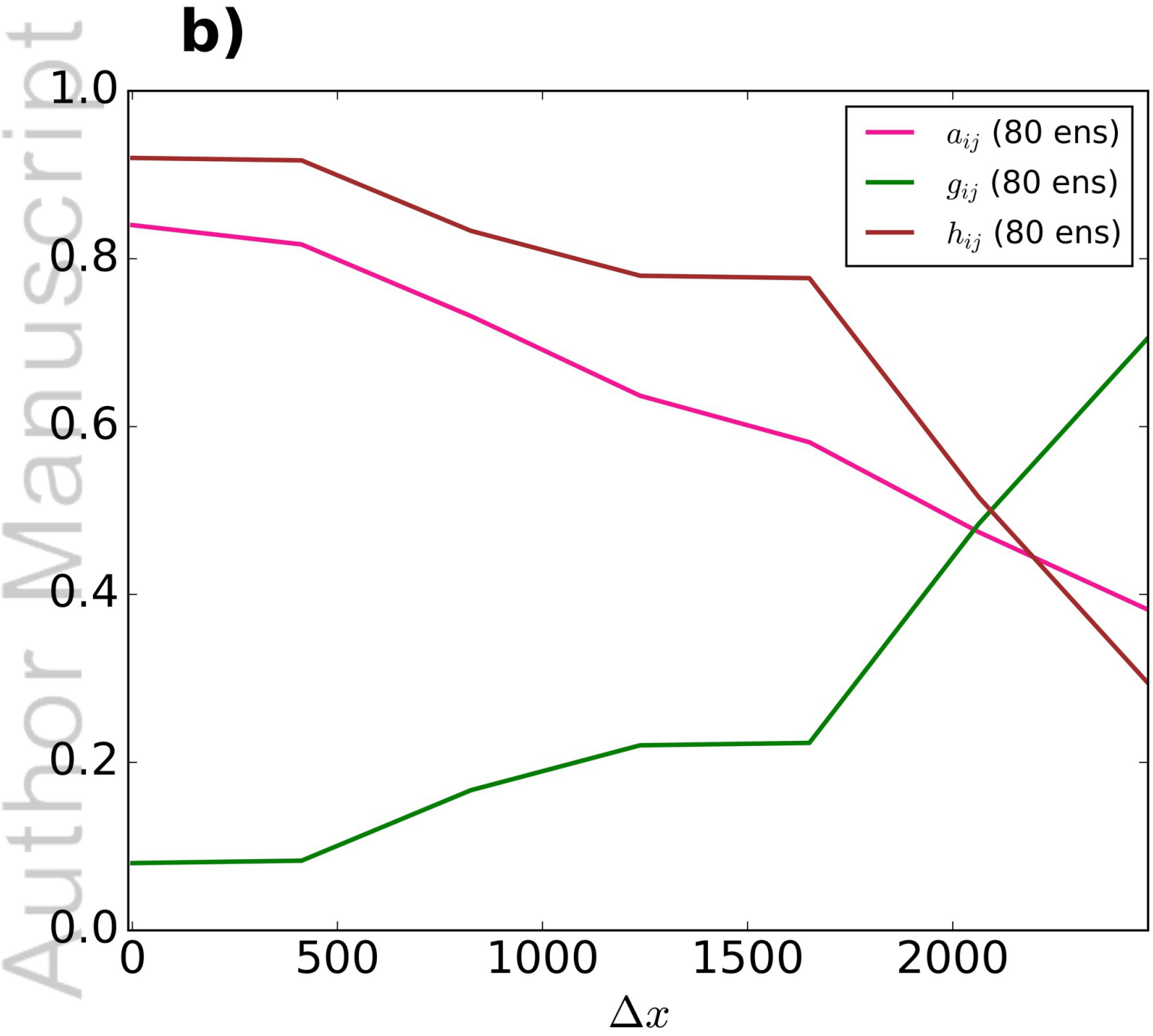
Figure 9. $\overline{RMSD}^{(t,\lambda,\phi)}\left(d_{ij}\right)$ *associated with the* $\widetilde{B_{ij}^{c}}$, $\left\langle\left(P_{ij}^{ens}\right)\right\rangle$, $\left\langle C_{ij}^{B}\left(P_{ij}^{ens}\right)\right\rangle$ *and the traditional Hybrid for an ensemble of 20 members.* $\overline{RMSD}^{(t,\lambda,\phi)}\left(d_{ij}\right)$ *refers to the square root of the average of the square differences, between the distinct covariance models and the Hybrid, over all the grid points with latitudes between 57.5 S and 57.5 N and also averaged over all the numerical simulation period.*



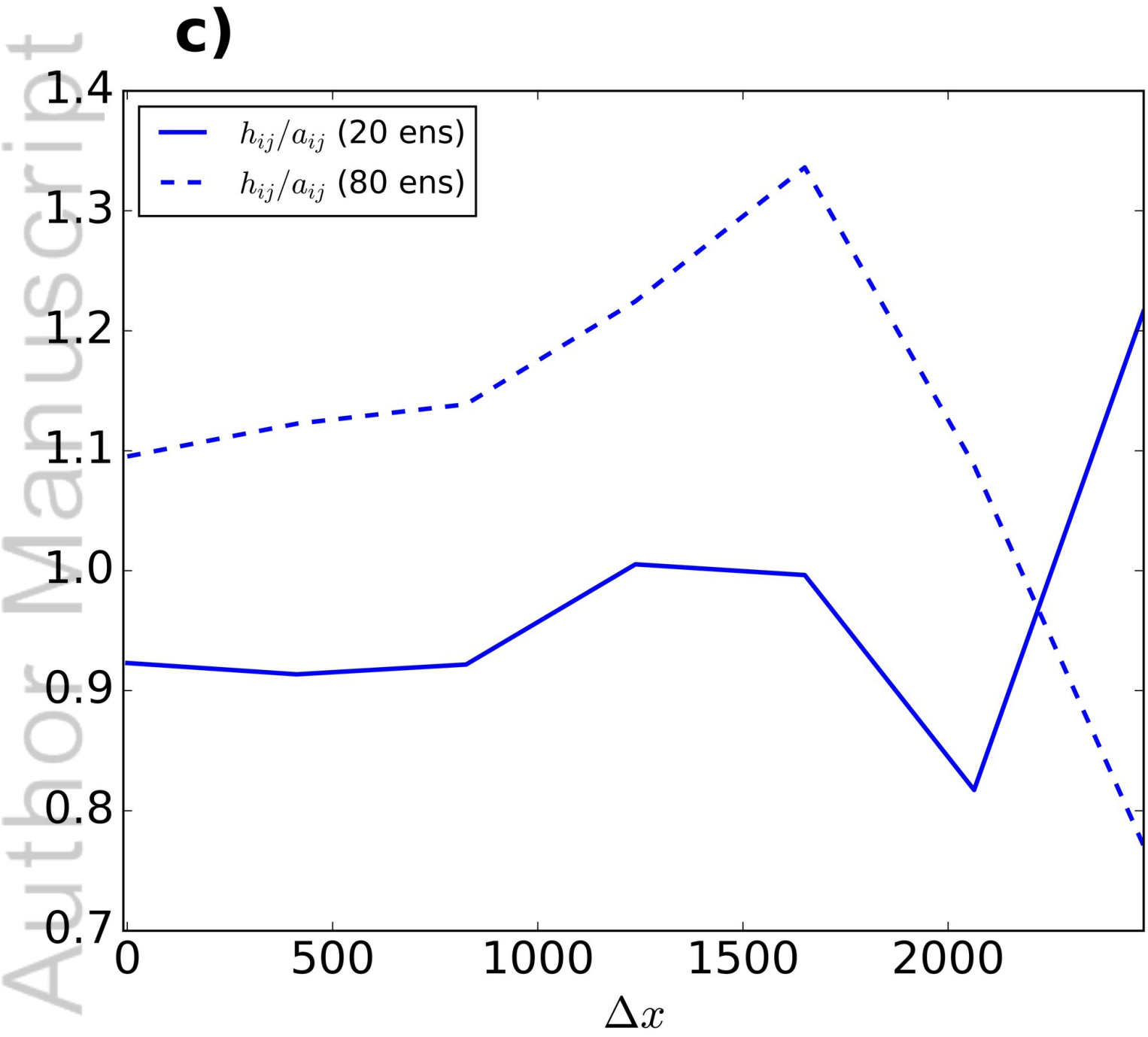
QJ_4008_Carrio_Fig4.jpg



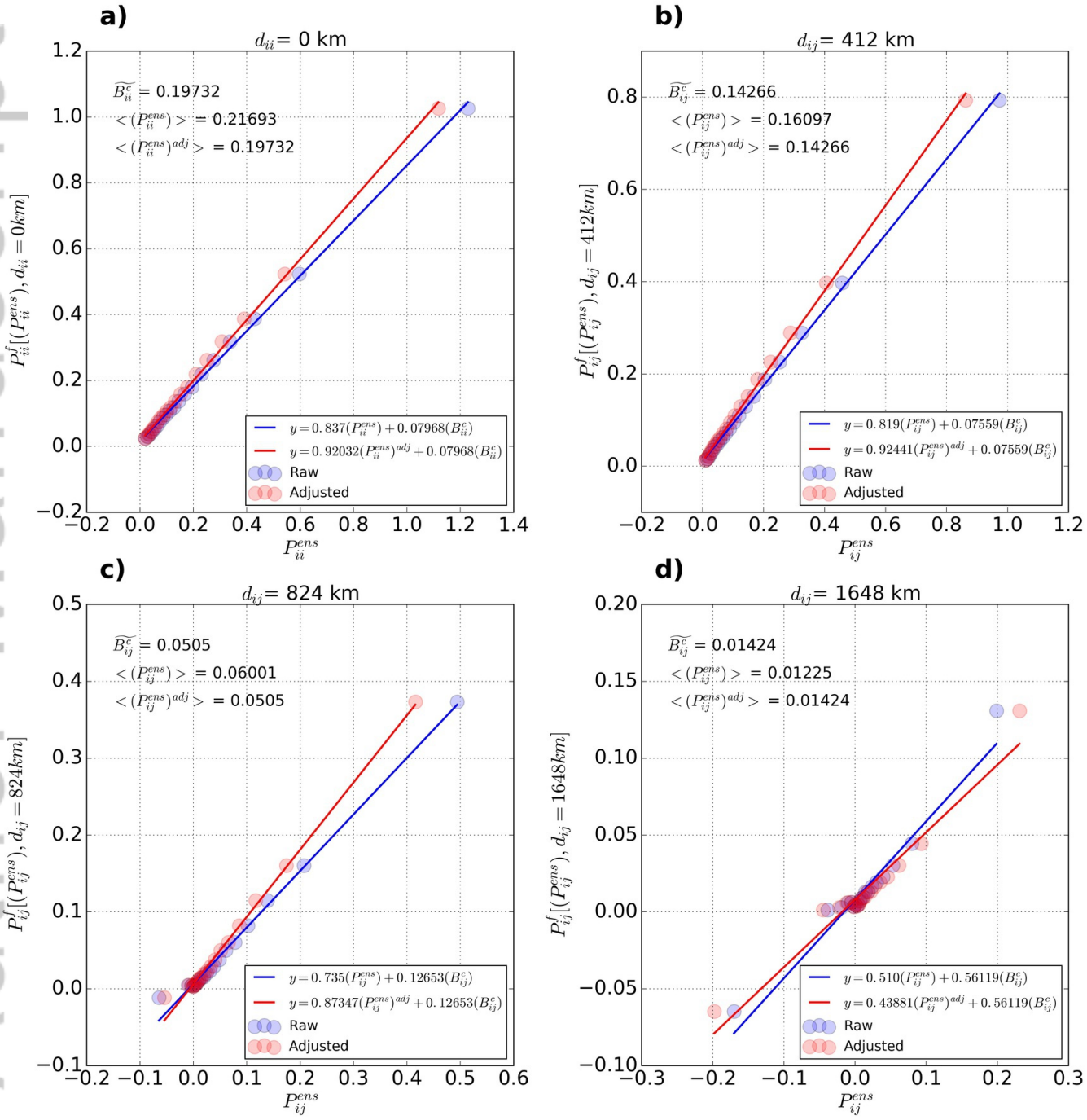
QJ_4008_Carrio_Fig5a.jpg



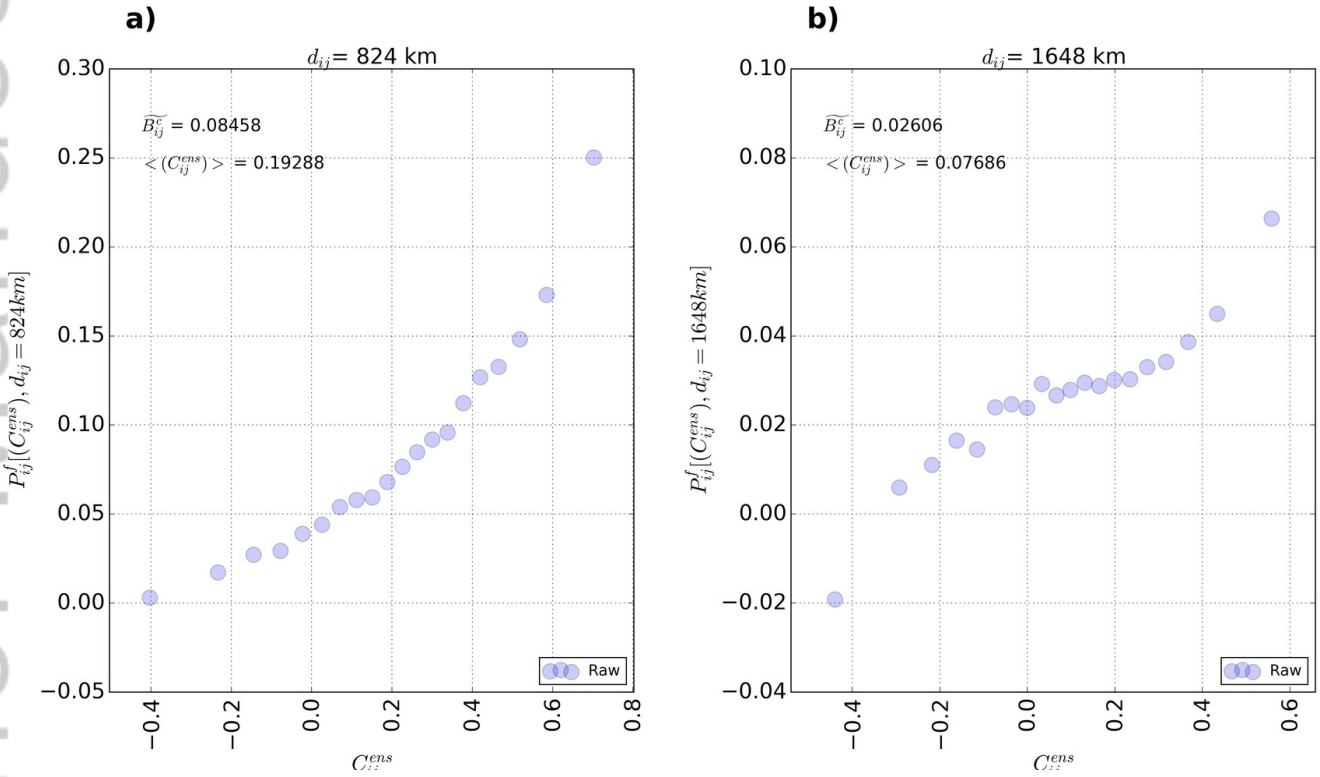
QJ_4008_Carrio_Fig5b.jpg



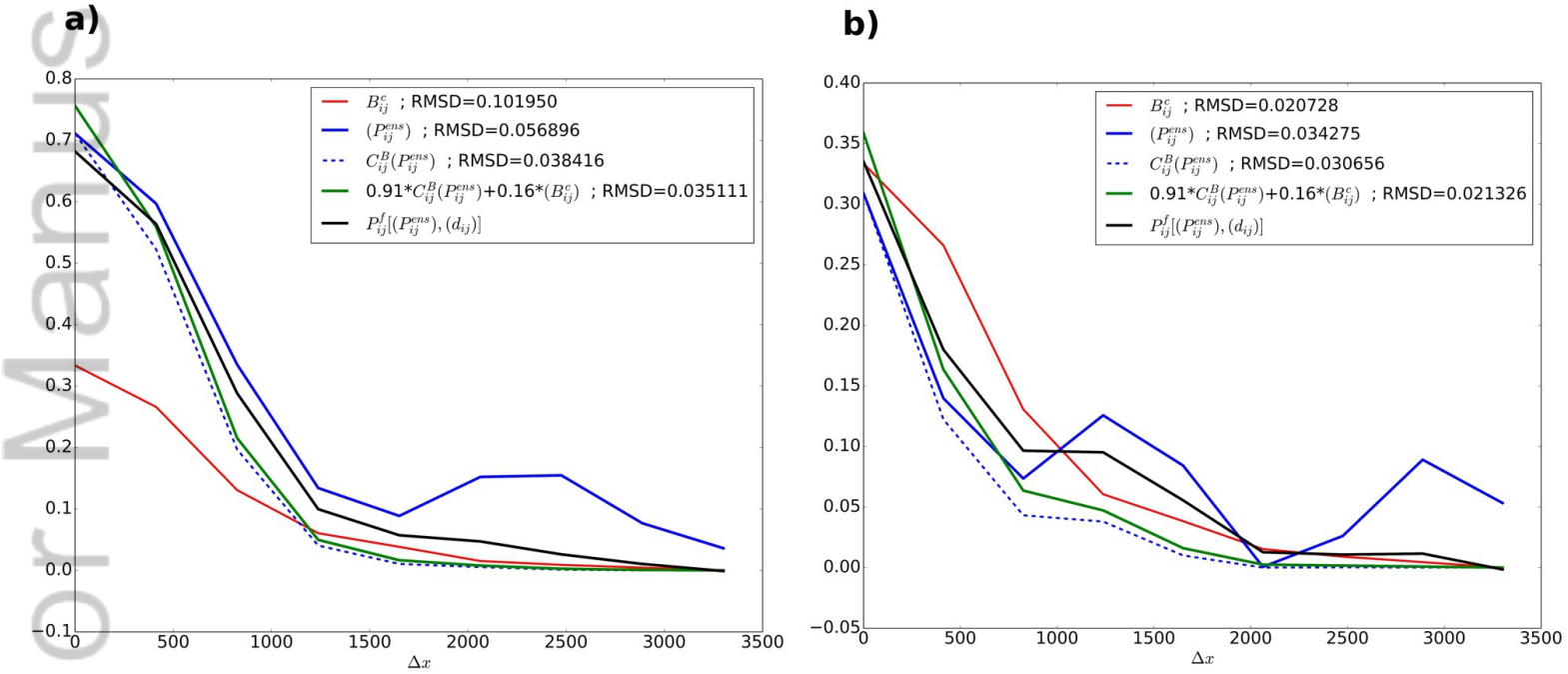
QJ_4008_Carrio_Fig5c.jpg



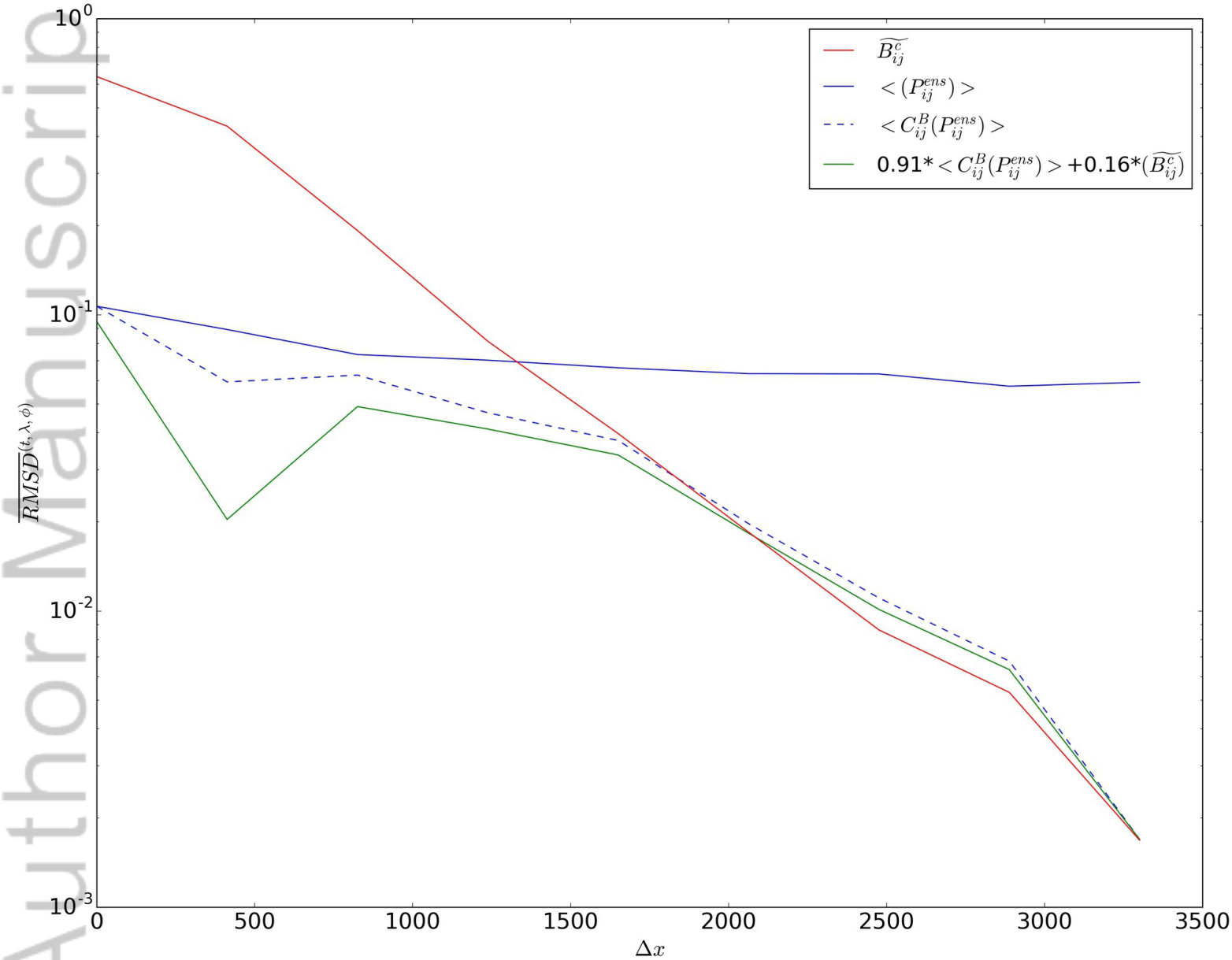
QJ_4008_Carrio_Fig6.jpg



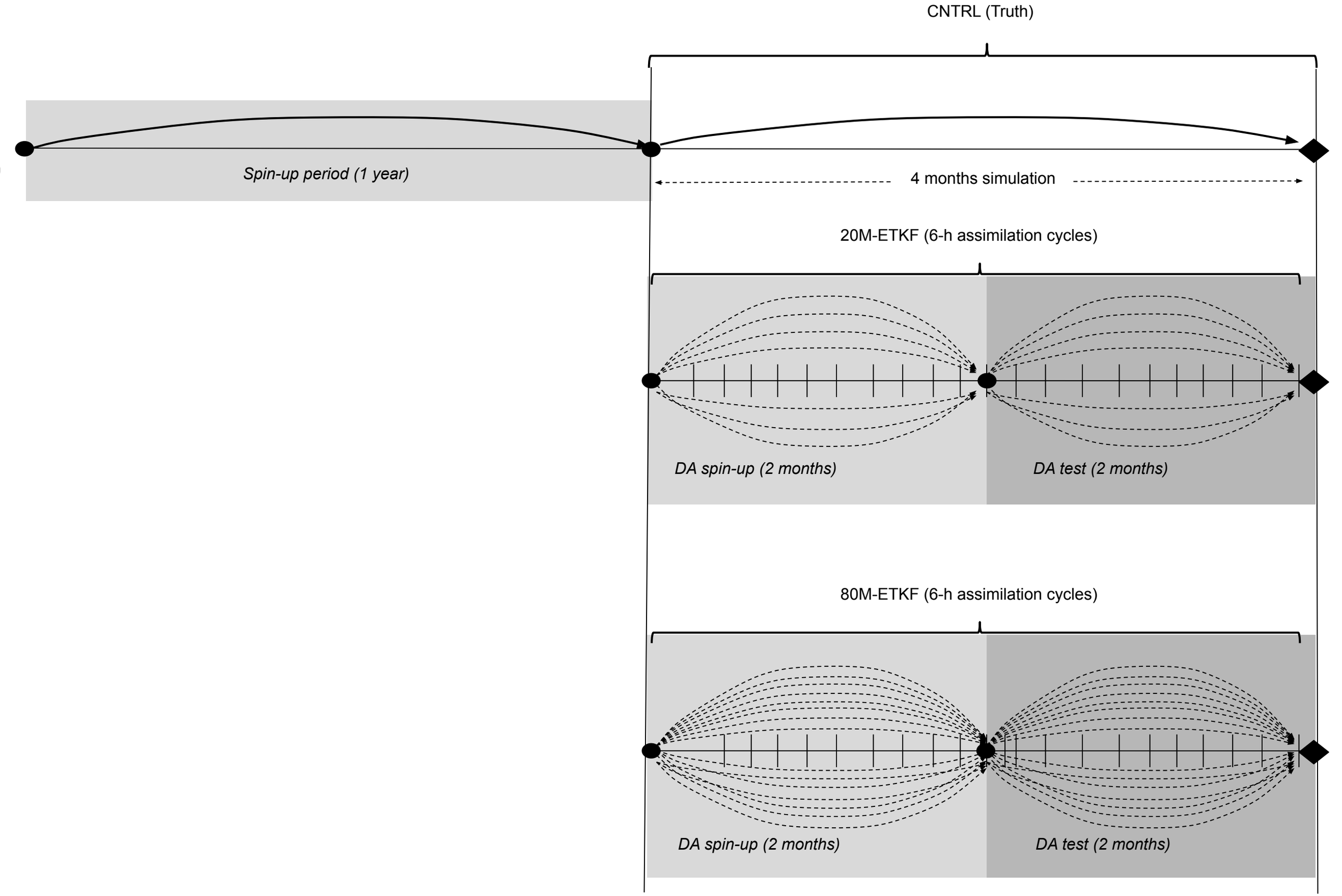
QJ_4008_Carrio_Fig7.jpg

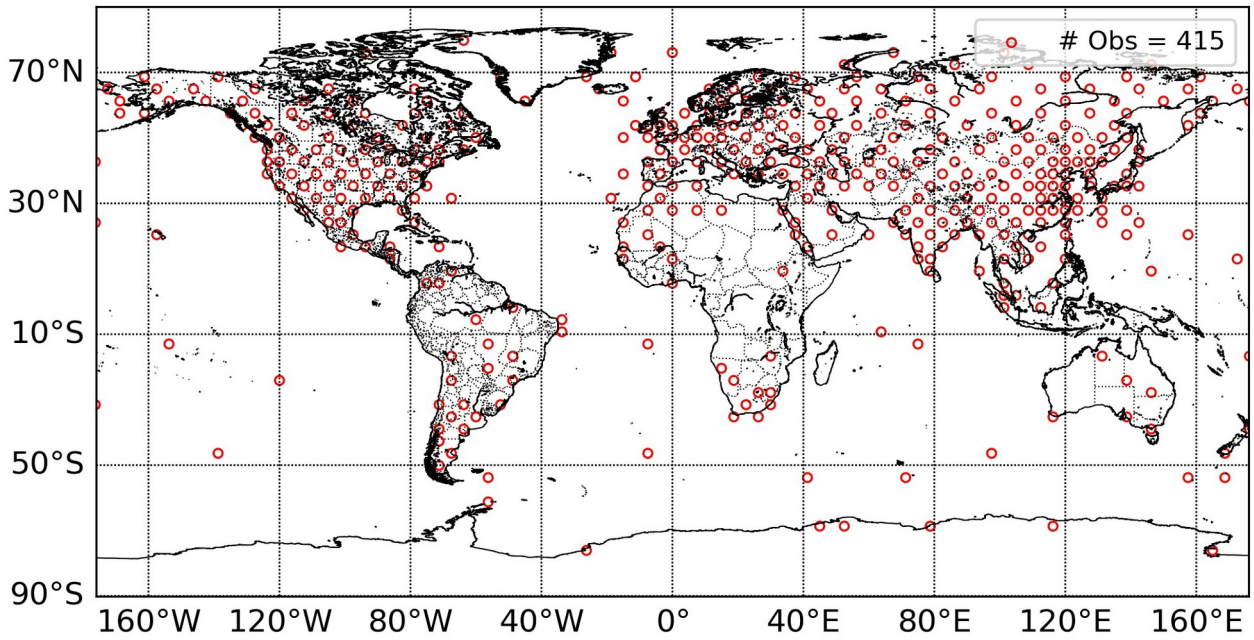


QJ_4008_Carrio_Fig8.jpg

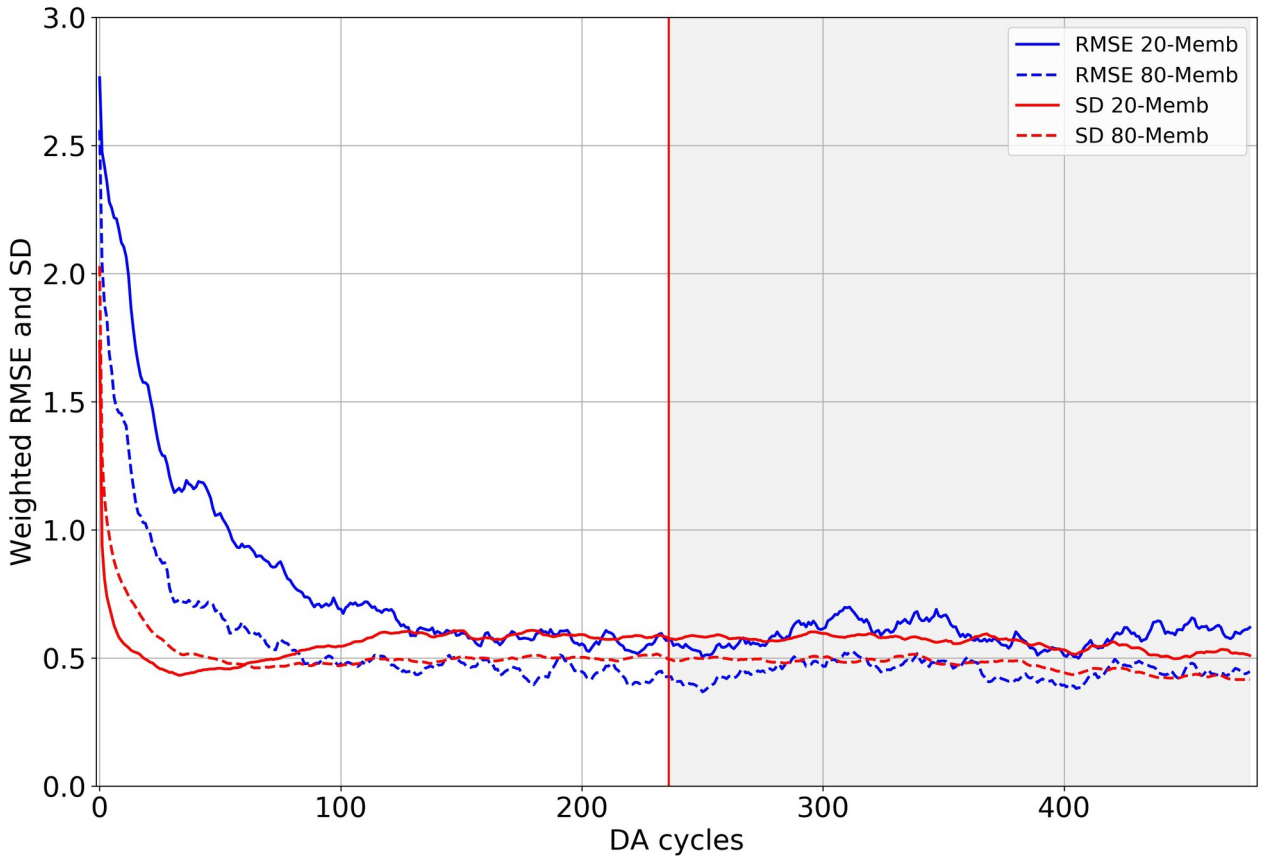


QJ_4008_Carrio_Fig9.jpg


CNTRL (Truth)
Spin-up period (1 year)
4 months simulation
20M-ETKF (6-h assimilation cycles)
DA spin-up (2 months)
DA test (2 months)
80M-ETKF (6-h assimilation cycles)
DA spin-up (2 months)
DA test (2 months)



QJ_4008_Figure_2.jpg



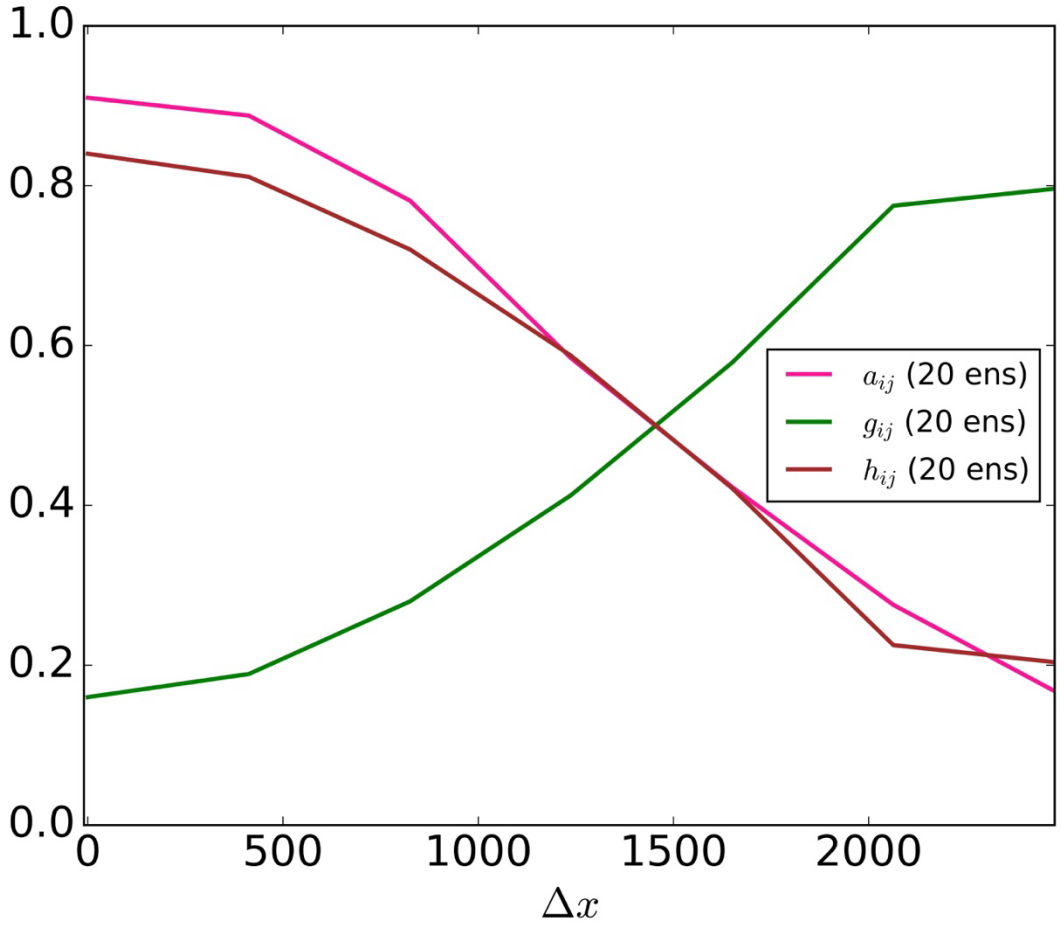
QJ_4008_Figure_3.jpg

Empirical determination of the covariance of forecast errors: an empirical justification and reformulation of Hybrid covariance models.

D. S. Carrió[a,,b,], C. H. Bishop[a,b], S. Kotsuki[c,d,e]*

[a]*School of Earth Sciences. The University of Melbourne, Parkville, Victoria, Australia*
[b]*ARC Centre of Excellence for Climate Extremes*
[c]*Center for Environmental Remote Sensing, Chiba, Kobe, Japan*
[d]*RIKEN Center for Computation*
[e]*PRESTO, Japan Science and Technology Agency, Chiba, Japan*

Table of Content



The introduction of Hybrid models into several forecasting centres has led to significant forecast improvements in the last years. This study empirically demonstrates that Hybrids provide a much better approximation to the true error covariances than error covariance models based on climatological error statistics and also to those based on localized ensemble-based dynamic error covariance estimates. The study also reveals and explains two fundamental deficiencies of current Hybrid error covariance formulations.