Referee Report

Manuscript egusphere-2026-2653

"Estimating cross-correlation parameters between forecast and observation errors within the ensemble transform Kalman filter with cross correlation (ETKFCC)"

Y. Kobayashi, S. Ohishi, and T. Miyoshi

Recommendation: Major revision.

The manuscript extends the innovation-statistics diagnostics of Desroziers et al. (2005) to the case in which the observation errors are cross-correlated with the forecast errors, and proposes an offline procedure for estimating the contamination factor and uncorrelated noise variance governing that cross-correlatio. The estimated parameters are then inserted into the ETKFCC formulation of Ohishi et al. (2025, "OKM25"). Feasibility is assessed through perfect-model twin experiments with the 40-variable Lorenz-96 model, comparing three configurations: the standard ETKF, the ETKFCC with estimated parameters, and the ETKFCC with prescribed true parameters.

General assessment

I have re-derived every equation in Sects. 2.2–2.3 and Appendices B and C independently. The algebra is correct throughout; I found no substantive derivation error, and the consistency claim between Eqs. (C11) and (C14) holds as stated. The estimation error decomposition in Eq. (40) is exact. The manuscript is therefore mathematically sound at the level of its stated derivations, it addresses a genuine practical obstacle to the deployment of OKM25, and the finding that the method remains useful when the underlying ETKF is only suboptimally tuned (Fig. 6) is the most operationally relevant result in the paper.

My concerns are of a different kind. The derived identities hold only under conditions that the experiments do not clearly meet. It is unclear which analysis covariance was actually substituted into half of the estimators. And the observation-generation design allows each filter to influence its own observations, which contaminates the headline comparison. None of these requires new theory: a clarifying subsection and two further experiments would resolve all three. I also suggest, in what follows, simpler explanations for several of the paper's own results.

Major comments

M1. The estimator implicitly requires the assimilating gain to be the optimal KFCC gain, and this is never stated. Equations (24) and (26) follow from Eqs. (B2)–(B3), which hold only for the optimal gain $\acute{\mathbf{K}}$. For a general gain, $\langle \boldsymbol{d}^{a-b}(\boldsymbol{d}^{o-b})^{\mathrm{T}} \rangle$ acquires a factor equal to the full innovation covariance rather than reducing to $\rho\mathbf{H}\mathbf{P}^{f}\mathbf{H}^{\mathrm{T}}(\mathbf{I}-\mathbf{A})^{\mathrm{T}}$. The statistics substituted into Eqs. (29), (30) and (34)–(36) are, however, taken from the *standard ETKF*, whose gain is not $\acute{\mathbf{K}}$.

In the scalar case the estimator is unbiased if and only if the assimilating gain equals $\acute{\mathbf{K}}$ exactly, and this is achievable by tuning a scalar R. The estimator therefore succeeds here because the ETKF was tuned to RMSE optimality, which in this idealised setting ($\mathbf{H}=\mathbf{I}$, uniform a, scalar R and ρ) drives the ETKF gain to essentially the KFCC gain. This, rather than ensemble-spread quality alone, is in my view the primary cause of the degradation of a^{est} away from the white hatches in Fig. 5a.

The point bears directly on the transferability claimed in the Summary: with heterogeneous observation types, a non-identity $\mathbf{H}$, and a single scalar inflation, the tuned ETKF gain

cannot equal $\acute{\mathbf{K}}$ at every location. Please state the requirement explicitly where Eqs. (24) and (26) are introduced, give the general-gain form, and revise the Summary accordingly.

M2. $\mathbf{H}\acute{\mathbf{P}}^a\mathbf{H}^{\mathrm{T}}$ is required by half the estimators but does not exist in the runs supplying the statistics. Equations (28), (30), (32), (33), (35) and (36) all require the analysis error covariance of the *cross-correlated* filter, yet the estimation uses standard ETKF output, which yields $\tilde{\mathbf{P}}^a$ and $\widehat{\sigma}^a$. The manuscript nowhere states what was actually substituted. This is a structural inconsistency, and I suspect it, rather than the error-cancellation argument of ll. 236–240, is the dominant reason why the estimators built on Eq. (30) and Eqs. (35)–(36) perform poorly. Please state precisely which quantity was used and discuss the inconsistency.

M3. The observation-generation design contains a feedback loop. At l. 166 the forecast error is computed online from the experiment being evaluated, so for $a > 0$ a more accurate filter generates less contaminated observations *for itself*. The three experiments do not assimilate the same observations, and the RMSE-ratio normalisation rescales this feedback rather than removing it; some fraction of the reported 5% is therefore a self-reinforcement artefact. Generating the contaminating forecast error from an independent reference run, held fixed across all three experiments, would fix this. I regard the correction as necessary before the quantitative improvement figures can be accepted.

M4. A short error analysis would replace three separate qualitative explanations. Propagating a forecast-spread error through Eqs. (29) and (34) shows that its effect on a^{est} is amplified by $1/MSE^f$, whereas its effect on r^{uc_est} is amplified only by $(1-a)^2 \leq 4$. Since $MSE^f \ll r^{uc} = 1$ throughout, a^{est} is one to two orders of magnitude more sensitive. Two lines of algebra thus account simultaneously for the contrast between Fig. 2a and Fig. 2b, for the statement at ll. 279–280 that r^{uc_est} is less sensitive to tuning, and for the complete failure at $a^{true} < 0$, where MSE^f is smallest. I recommend replacing the qualitative discussion of ll. 236–240 with this argument.

M5. The non-uniform estimators are derived and never tested. Equations (27), (28) and (31)–(33) treat spatially varying a_i and r_i^{uc}, but all experiments use uniform parameters and the general case is deferred (ll. 325–328). However, the stated motivation is precisely a setting in which contamination varies with observation density. One experiment with a spatially varying a_i using Eqs. (27) and (31) would close this gap. I note in passing that a diagonal $\mathbf{A}$ is a modelling choice, not a necessity: Eq. (24) determines the full matrix whenever $\mathbf{H}\mathbf{P}^f\mathbf{H}^{\mathrm{T}}$ is invertible, and this is worth remarking on.

Minor comments

m1. Equation (34) is nonlinear in the innovation statistics, so replacing expectations by time averages introduces a Jensen-type bias that does not vanish with sample length when $\mathbf{H}\mathbf{P}^f\mathbf{H}^{\mathrm{T}}$ varies in time. Please comment, and ideally quantify.

m2. The $a \to 1$ limit is an identifiability degeneracy, not a tuning failure: at $a = 1$ the innovation reduces to pure noise and carries no information about the forecast error, so no innovation-based estimator can recover a. Lines 248–250 present the divergence at $a^{true} = 0.9$ as a consequence of a^{est} being near unity, which reverses the causality.

m3. The failure at $a^{true} < 0$ is worth saying so: a^{est} is small, the ETKFCC reverts to the ETKF, and Fig. 4 confirms no loss. Since the sign of a is unknown in practice, a failure mode that degrades to the baseline rather than below it is a positive property, not merely an unrealistic case to be dismissed.

m4. The correlation coefficient depends on MSE^f and therefore differs between the three experiments, yet Fig. 4b plots three curves against a single axis. Please state which experiment defines that axis and give the formula in the text, since the headline claim ("0.2–0.7") rests entirely on it.

m5. Figure 2c is dominated by the blow-up at $a^{true} \geq 0.7$, so the spread errors for $a^{true} < 0$ that ll. 223–225 invoke are indistinguishable from zero. A relative ordinate ($\delta\widehat{\sigma}^2/MSE^f$) or a symmetric log axis would make the argument legible.

m6. No sample-size sensitivity is shown. A convergence curve of a^{est} and r^{uc_est} against estimation-window length would be valuable for any operational claim.

m7. Equations (10) and (13) require the symmetric square root and a mean-preserving sign convention for the transform to be unique. Please state this.

Technical corrections

T1. l. 110: "$\mathbf{C} \cong \rho\mathbf{A}\mathbf{H}\mathbf{P}^f\mathbf{H}^{\mathrm{T}}$" is the transpose of the intended quantity and is dimensionally inconsistent.

T2. Table A5: $\mathbf{H}$ is listed as $\mathbb{R}^{n\times p}$; it is $\mathbb{R}^{p\times n}$.

T3. Table A1: a and r are given identical definitions.

T4. Abstract and l. 33: the OKM25 range (0.2–0.8) and the present range (0.2–0.7) sit close together without comment; a clarifying clause would help.

Closing

The novelty relative to OKM25 is moderate, but the extension is correctly derived and removes the main obstacle to applying that earlier work. The robustness result of Fig. 6 is undersold relative to the headline improvement figure. I would be glad to see a revised version.