Response to Referee #1
for "Elucidating the performance of data assimilation neural networks for chaotic dynamics"

Egusphere-2026-245

Marc Bocquet, Tobias S. Finn, Sibo Cheng, Alban Farchi

12 June 2026

This paper tries to elucidate the reasons of the impressive results obtained by Data Assimilation Networks (DANs) in Bocquet et al. 2024 (Boc24 in the manuscript), by leveraging explainability techniques relying on the sensitivity of their Jacobian matrices. In particular, they provide a satisfactory proof that the scalability of the method to "unseen" higher dimensional model versions is due to its focus on local patterns.

The manuscript is well-written and interesting to read, and achieve the stated goal, providing a much needed explainability framework, a feature usually absent from most works using machine learning (ML). I recommend therefore its publication once the following comments have been addressed.

We sincerely thank the Reviewer for the positive evaluation of the manuscript and for the constructive comments. In the following, we discuss the raised concerns and what we have changed or revised in the manuscript. The revised manuscript and a pdf document highlighting the differences between the original and the revised manuscript will be provided.

(1) My main feeling is that the article go a bridge too far when it comes to discarding the usefulness of ensembles. For example lines 30-31: 'This result challenges the long-standing assumption that explicit ensemble representations are indispensable to estimate flow-dependent uncertainties in chaotic systems.' This may be true for Data Assimilation (DA), but it is not sure that this holds with respect to other analysis, where ensemble representations (or probabilities) might still present some usefulness.

We agree with this important qualification. Our intention was not to argue that ensembles are generally unnecessary, nor that probabilistic representations can be dispensed with in forecasting, uncertainty quantification, etc. Our result is narrower: in the perfect-model filtering experiments considered here, the point-estimation accuracy of a learned analysis operator can match ensemble-based baselines even when the operator is driven by a single forecast state. Actually, in parallel to using DAN, we keep using ensemble DA methods for DA studies, either as main DA methods or as baseline against ML-enhanced methods. Note however that the sentence that you quote does not imply that the ensembles have become useless, but that alternatives to ensembles for UQ may exist; and it is true that we have only shown this in the context of DA.

Added subsection in the main discussion section:

Ensembles, uncertainty quantification, and multiplicative ergodic theorem

Our results show that, in the perfect-model filtering setting considered here, an explicit forecast ensemble is not strictly required for a learned analysis operator to extract the flow-dependent information needed to achieve EnKF-like point-estimation accuracy. This conclusion should not be interpreted as a dismissal of ensemble representations in general. Ensembles remain essential for probabilistic forecasting, uncertainty quantification, the diagnosis of model and observation errors, smoothing, risk-sensitive applications, and regimes in which the posterior distribution is strongly non-Gaussian or multi-modal.

Moreover, the deterministic analysis operator $\mathbf{a}_{\boldsymbol{\theta}}$ could itself be used as a building block for ensemble generation. For instance, applying $\mathbf{a}_{\boldsymbol{\theta}}$ to perturbed innovations would produce a set of analysis states, in a way that is consistent with the local, innovation-dependent interpretation used in, e.g., Appendix I to obtain the linear regression approximation of $\mathbf{a}_{\boldsymbol{\theta}}$ with respect to $\boldsymbol{\zeta}$. Such an

analysis ensemble could then be propagated by the forecast model to estimate forecast uncertainty. We have not pursued this probabilistic use of DAN in the present study, since our focus was on point-estimation accuracy, but it represents a natural direction for future work.

(2) The authors should also comment on the fact that besides DA, the determination of flow-dependent uncertainties using ML has already been studied, with variable success. This raises the question of why it works so well here. One could conjecture that is due to the information on uncertainties (and instabilities) needed by the DA processes are actually suitable for its inference using ML, while determining actual precises quantities such as the Covariant Lyapunov Vectors (CLVs) and Lyapunov exponents is a more challenging task. The fact that a Multiplicative Ergodic Theorem (MET) exists for the underlying systems is clear, but this provides a mapping between the states of a system and its CLVs, it doesn't mean that this mapping between the CLVs at a given time can be determined alone from the state at the same time. For example, the Ginelli algorithm combines a forward and a backward pass, which take some "time" to converge (see F. Noethen studies to have an idea on this). Therefore it may explain why ML sometimes struggle to learn this mapping. Here, for DA, it works very well, and the DAN seems to be able to learn what is useful, even with just one member, but somehow this is an easier task than learning the MET mapping. In the end, it is probably connected to the fact that CLVs are non-local (and therefore dimensional scalability of algorithms computing them is not clear), contrary to the mapping between the forecast state and the analysis error covariance, as shown by your work.

We believe that our manuscript did not go in any way against your comment. We take your comment as an invitation to discuss this very interesting point.

By *The fact that a MET exists for the underlying systems is clear, but this provides a mapping between the states of a system and its CLVs, it doesn't mean that this mapping between the CLVs at a given time can be determined alone from the state at the same time.*, we guess that what you have in mind is: *MET ensures that, for almost every state on the attractor, the Oseledets subspaces/CLVs are state-dependent measurable objects. However, MET alone does not imply that this state-to-CLV map is regular, local, or practically recoverable from instantaneous state samples without information about the tangent dynamics along trajectories.*

In that sense, we agree with you that the existence of the Oseledets splitting does not by itself establish that the map from state to CLVs is smooth, continuous, or easily learnable from finite data. However, for the deterministic systems considered here, MET does imply that the CLV/Oseledets subspaces are measurable functions of the full state, for almost every state on the attractor. We have revised the text to make clear that our claim concerns this almost-everywhere state dependence, not the stronger assertion that the CLVs can be obtained by a simple local diagnostic from the instantaneous state alone.

One should also keep in mind that the DA process as a dynamical system is asymptotically stable as opposed to the forecast dynamics it is based upon. That may be critical in explaining why learning such mapping is easier, as opposed to, e.g., learning a mapping from the state to the CLVs.

Last part of the added subsection the main discussion section: At a more theoretical level, the success of DAN may be facilitated by the existence, in the present setting, of a sufficiently regular and local dependence of the relevant flow-dependent analysis correction on the forecast state. This should not be taken for granted in general. For example, the multiplicative ergodic theorem ensures that covariant Lyapunov subspaces are defined as measurable functions of the state, for almost every state on the attractor, but it does not guarantee that this state-to-Oseledets-splitting map is continuous, local, or easily learnable from finite data. Other applications beyond DA, such as learning covariant Lyapunov vectors directly from the state, may therefore involve maps with poorer regularity or locality, making them substantially more difficult to learn and possibly less scalable. This question is left for future work.

(3) Lines 67-68: The analysis is an estimator of the conditional probability density function, or an estimator of its first moment ? Please clarify.

Thank you for pointing out this imprecision. The analysis state is a point estimator of the hidden state, obtained from the conditional distribution. Depending on the filtering formulation

and loss, it can be interpreted as a posterior mean, a MAP estimate, or another summary of the conditional distribution. In the present deterministic least-squares training, it is best described as a point estimator approximating the conditional mean under the adopted loss. We have revised the sentence accordingly.

Changed text: A filtering DA scheme estimates the hidden state $\mathbf{x}_k^{\mathrm{t}}$ at t_k from the observations available up to that time. The analysis $\mathbf{x}_k^{\mathrm{a}}$ is a point estimator, such as the posterior mean or maximum a posteriori estimate, of the conditional probability density function $p(\mathbf{x}_k^{\mathrm{t}}|\mathbf{y}_k, \mathbf{y}_{k-1}, \ldots, \mathbf{y}_1)$. A *sequential* filtering DA scheme infers $\mathbf{x}_k^{\mathrm{a}}$ from $\mathbf{y}_k$ and from background information about the state at t_{k-1} (possibly brought forward to t_k using $\mathcal{M}$).

(4) Does the argument on translational invariance holds also because CNNs are know to be shift invariant, meaning the method used here would not be applicable to DNN without this invariance ?

To be precise, convolutional neural networks are translation equivariant, not strictly translation invariant, until an invariant reduction such as time-averaging is applied. The translational symmetry is first a property of the L96 model and of the homogeneous observation configuration; the convolutional architecture then encodes this equivariance by weight sharing and makes the learning problem much more data-efficient. A fully connected DNN could, in principle, learn the same map if supplied with enough data and suitable augmentation, but it would not encode the symmetry, would have many more parameters, and would not naturally transfer to a different value of N_{x}. Thus the method is not mathematically inapplicable without a CNN, but the demonstrated scalability relies on using an architecture compatible with the symmetry and locality of the problem.

(5) Basically the authors show that the CNNs DANs can be generalized to "unseen" cases (i.e. unseen higher dimensional version of the model at hand), because actually they focus on a subset of features (i.e. the local features) common to most of the model versions. Does that means that in the case of models encountering a dramatic change in its local properties (such as a change in the logarithmic slope of the energy spectrum for example) when the dimensionality is increased, DAN generalization would not work ?

Yes, this is an important limitation. The generalisation is expected to work only when the local dynamics and local forecast-error structures seen during training remain representative of those encountered at the new resolution or model size. If increasing the dimension changes the local physics, the effective energy spectrum, the balance relations, or the observation-induced error structures, then a DAN trained at the original dimension may fail and would likely require retraining, conditioning on the new regime, or a multiscale architecture. We had tested this on a (continuous) Kuramoto-Sivashinsky model with two distinct spectral resolutions over the same domain. The DAN trained on one resolution would, out of the box, fail when tested on the other resolution, as expected. We have added this caveat to the manuscript.

Changed text: This neat scalability has limitations. If increasing the dimension changed the local statistics or local dynamical balances of the model, for example through a substantial change of spectral slope or local instability mechanisms, a DAN trained at the smaller dimension should not be expected to generalise without fine-tuning, additional training, or architectural adaptation.

(6) The manuscript is not self-contained enough, lots of details are simply mentioned as being from Boc24 and the reader has to go there to understand DANs details. Therefore, if in the future it becomes (more) difficult to find Boc24 for any reason, this manuscript would cease to be understandable. Could the authors incorporate a reasonable amount of the Boc24 setup description in this paper as well? Like for example a figure of the CNNs DAN schematics.

We agree and we have created a new appendix (now Appendix A), referred from the methodological section, to provide the details on the architecture, and we have supplemented the appendix on the sensitivity to the training parameters (now Appendix B) with a figure describing how the

data are organised and fed to the training scheme.

(7) There is a typo in the x label of fig 7.

The transpose does appear when using the original pdf and using any pdf viewer we have access to. Figure 6 is also impacted (background grid). We have checked that all the required fonts are embedded in both the standalone figure and in the original manuscript pdf. Hence, the missing transpose seems to be due to Copernicus' processing of the figures.

(8) Line 131: 'memorises' is maybe a bit too much anthropomorphic here.

As much as we would like to agree with you, *to memorise* is also commonly used for machines. And it is probably easier to pinpoint the memory footprint in a CNN than in a human brain, notably in the filters of the convolutional layers. Some neural networks, such as echo state networks are explicitly based on the ability to memorise patterns.

Response to Referee #2, Patrick N. Raanes for "Elucidating the performance of data assimilation neural networks for chaotic dynamics"

Egusphere-2026-245

Marc Bocquet, Tobias S. Finn, Sibo Cheng, Alban Farchi

12 June 2026

The Reviewer finds the manuscript to be of high quality, with impressive results, and recommends publication after minor changes. The comments mainly ask for clearer definitions, less circular explanations, improved notation, more explicit links to existing data-assimilation results, and greater clarity in the interpretation of the linear-in- ζ and nonlinear DANs.

We sincerely thank the Reviewer for the careful and constructive report, and in particular for the detailed chronological comments. In the following, we discuss the raised concerns and what we have changed or revised in the manuscript. The revised manuscript and a pdf document highlighting the differences between the original and the revised manuscript will be provided.

(1) Lines 2–5: The distinction between learning a direct mapping from forecasts and observations to increments and learning the analysis in the DAN setting was not clear.

We agree that the distinction was not clear enough in the abstract. At the risk of a long exposition in the abstract, we are now much more explicit so as to separate the two learning paradigms. In particular DANs are now introduced as the second paradigm hopefully reducing the risk of confusing the reader.

Changed text (first part of the abstract): In supervised data assimilation machine learning emulation, the training data contain targets produced by an existing data assimilation scheme, such as analysis increments. By contrast, *data assimilation networks* were recently proposed to learn the analysis operator while they are embedded in the forecast–analysis cycle: their only targets are the true trajectory and the observations thereof. They are therefore trained to produce a stable and accurate sequential estimator, rather than to reproduce the output of a prescribed data assimilation algorithm. Conceptually more fundamental, yet computationally more challenging, such learned data assimilation scheme was shown to achieve accuracy comparable to that of the ensemble Kalman filter when applied to low-order chaotic dynamics. Strikingly, the same accuracy can be reached with a single state forecast instead of an ensemble, hence bypassing the need to explicitly represent forecast uncertainty.

(2) Equation (5): The role of the linear form was unclear and looked circular. Does it define the expansion, P, or its computation/approximation?

We assume that the source of confusion was in the precise definition of what $\mathbf{P}^{\mathrm{a}}$ stands for in Eq. (5), due to an unclear indeed circular wording. Equation (5) is the generic first order expansion of $\mathbf{P}^{\mathrm{a}}$ in $\boldsymbol{\zeta}$. When compared to the theoretical Kalman update and its linearisation in $\boldsymbol{\zeta}$, the linear operator acting on $\boldsymbol{\zeta}$ should be an analysis error covariance matrix, which can presently be attached to the analysis of DAN. We have changed the text around Eq. (5) to clarify.

Changed text: To unveil the mechanisms leveraged by the learned analysis operator to achieve this accuracy, Boc24 performed a linearisation of the operator $\mathbf{a}_{\boldsymbol{\theta}}$ in the projected innovation $\boldsymbol{\zeta}$:

$$\mathbf{a}_{\boldsymbol{\theta}}(\mathbf{x}, \boldsymbol{\zeta}) \approx \mathbf{P}^{\mathrm{a}}(\mathbf{x})\boldsymbol{\zeta}, \tag{1}$$

where $\mathbf{P}^{\mathrm{a}}(\mathbf{x})$ is the resulting linear operator that acts on $\boldsymbol{\zeta}$. Comparing with the linearisation of the classical Kalman update, it must coincide with the analysis error covariance matrix which could be associated by formal analogy to the learned analysis operator $\mathbf{a}_{\boldsymbol{\theta}}$, hence the notation.

(3) Line 112: The definition of ζ should be more prominent.

This has been corrected: we now introduce $\boldsymbol{\zeta}_k$ immediately after the DAN analysis equation and consistently refers to it as the projected innovation.

Changed text: For compactness, we define the projected innovation

$$\boldsymbol{\zeta}_k \triangleq \mathbf{H}_k^{\intercal}\mathbf{R}_k^{-1}\boldsymbol{\delta}_k = \mathbf{H}_k^{\intercal}\mathbf{R}_k^{-1}\left(\mathbf{y}_k - \mathbf{H}_k(\mathbf{x}_k^{\mathrm{f}})\right). \tag{2}$$

The DAN analysis operator thus receives two fields in state space, $\mathbf{x}_k^{\mathrm{f}}$ and $\boldsymbol{\zeta}_k$.

(4) Line 116: Replace "hence" by "implying".

Corrected.

(5) Line 166: Δt had not yet been specified, reducing the usefulness of the stated RMSE values.

Indeed. $\Delta_t = 0.05$ is now part of the reference setup specification (defined a couple of paragraphs earlier).

Changed text: ... The time-step between observation batches and updates is set to $\Delta_t = t_{k+1} - t_k = 0.05$ for all experiments with the exception of Sect. 4.1. This will be our reference DA setup.

(6) Line 175: The assumption seems tied to quasi-linearity rather than the premises stated above, and "close to optimal" was not defined.

We agree. We have rewritten this passage to make explicit that the statement holds under a Gaussian/quasi-linear approximation. By "close to optimal" we meant that assuming Gaussianity/quasi-linearity, the analysis is given by the MAP estimate, which is also the posterior mean in the Gaussian/linear case. The wording has been made precise.

Changed text: Under the additional Gaussian and quasi-linear assumptions used in this subsection, a *close to optimal* analysis associated with the Kalman update is the posterior mean which coincides with the maximum a posteriori of the conditional pdf, hence by the minimum of the cost function associated to the analysis.

(7) Equation (8): Display $x^{\mathrm{a}} = x^{\mathrm{f}} + P^{\mathrm{a}}\zeta$ nearby and relate it to Eq. (2).

We have followed your suggestion.

Changed text: As a consequence, $\mathcal{J}$ is quadratic in $\mathbf{x}$, strictly convex, and its minimum argument is (Daley, 1991)

$$\mathbf{x}^{\mathrm{a}} = \mathbf{x}^{\mathrm{f}} + \left(\mathbf{H}^{\intercal}\mathbf{R}^{-1}\mathbf{H} + \left(\mathbf{P}^{\mathrm{f}}(\mathbf{x}^{\mathrm{f}})\right)^{-1}\right)^{-1}\mathbf{H}^{\intercal}\mathbf{R}^{-1}\left(\mathbf{y} - \mathbf{H}\mathbf{x}^{\mathrm{f}}\right) \tag{3a}$$

$$= \mathbf{x}^{\mathrm{f}} + \mathbf{P}^{\mathrm{a}}(\mathbf{x}^{\mathrm{f}})\boldsymbol{\zeta}, \tag{3b}$$

where the projected innovation $\boldsymbol{\zeta}$ has been defined by Eq. (2). Equation (3b) matches the linearisation Eq. (1).

(8) Line 185: The conclusion of the Gaussian subsection was unclear and seemed circular.

We have clarified that the subsection is not a proof that DAN is an EnKF. It only establishes that, under standard Gaussian/quasi-linear DA assumptions, the linearisation of a successful DAN with respect to $\boldsymbol{\zeta}$ can be interpreted as an effective Kalman-like covariance or gain. The numerical tests then assess whether this diagnostic is useful, and when it might break.

Changed text: The purpose of this derivation is to show that, under Gaussian/quasi-linear assumptions, DAN is likely to solve the optimisation problem Eq. (7), its update follows Eq. (3b), and that its derivative plays the role of an effective analysis error covariance matrix $\mathbf{P}^{\mathrm{a}}(\mathbf{x}^{\mathrm{f}})$. Whether this local interpretation is sufficient is then tested numerically.

(9) Line 198: The phrase "linear-in-ζ" needed a clear definition.

We have followed your suggestion.

Changed text: ... Details on our scalable implementation of Eq. (3b), for such a DAN linear in the projected innovations (later on called linear-in-$\boldsymbol{\zeta}$ DAN), can be found in Appendix C. ...

(10) Section 3.1.3: The theoretical arguments did not fit together clearly, and the last sentence was too strong.

We have re-arranged and practically rewritten Sect. 3.1.3, and toned down the conclusive sentence.

Changed text: Let us focus on Sakov (2025), whose algorithmic constructions targeted at estimating $\mathbf{P}^{\mathrm{f}}(\mathbf{x}^{\mathrm{f}})$ are especially transparent and whose numerical tests were carried out in the reference setup described in Sect. 3.1. Their Algorithm A1 proceeds as follows: (i) the state $\mathbf{x}^{\mathrm{f}}$ is backtracked by T time steps; (ii) the tangent linear model, evaluated along the resulting state trajectory from t_{-T} to t_0, is applied to a matrix of state perturbations $\varepsilon\mathbf{I}_{\mathrm{x}}$; and (iii) the resulting perturbations at time t_0 are used to estimate $\mathbf{P}^{\mathrm{f}}(\mathbf{x}^{\mathrm{f}})$. This procedure is closely related to the Assimilation in the Unstable Space (AUS) methods (Palatella et al., 2013; Carrassi et al., 2022), since for sufficiently large T the output of the tangent linear model at t_0 provides a square-root factor of the backward Lyapunov vectors. But it is already sufficient to yield a competitive analysis relying on the errors-of-the-day estimated through $\mathbf{P}^{\mathrm{f}}(\mathbf{x}^{\mathrm{f}})$ which notably relies on a single state.

Algorithm A2 of Sakov (2025) also backtracks the state $\mathbf{x}^{\mathrm{f}}$ by T time steps but differs in the subsequent step: instead of propagating perturbations, it applies a (square-root) Kalman filter to an initial covariance matrix $\varepsilon^2\mathbf{I}_{\mathrm{x}}$ from t_{-T} to t_0, yielding a more refined estimate of $\mathbf{P}^{\mathrm{f}}(\mathbf{x}^{\mathrm{f}})$ at t_0. This yields an even more accurate DA algorithm on par with a well tuned EnKF, again relying on $\mathbf{P}^{\mathrm{f}}(\mathbf{x}^{\mathrm{f}})$ assessed from a single state.

Bocquet et al. (2017); Bocquet and Carrassi (2017) have laid the theoretical grounds to understand why these principled demonstrations are successful. Bocquet et al. (2017) focused on a degenerate Kalman filter, equivalent to an EnKF in Gaussian and quasi-linear conditions. In this setting, the covariance evolution is formally decoupled from the state update, with the important caveat that the forecast error covariance $\mathbf{P}^{\mathrm{f}}$ depends on the state $\mathbf{x}^{\mathrm{f}}$ at time t_0. Indeed, Bocquet et al. (2017) showed that, asymptotically (i.e. for large T), $\mathbf{P}^{\mathrm{f}}$ depends on the system dynamics only through the Lyapunov vectors. By the MET, the corresponding Oseledets subspaces are measurable functions of the state $\mathbf{x}^{\mathrm{f}}$ at t_0 (for almost every state on the attractor). Hence, following Bocquet et al. (2017), Algorithm A2 by Sakov (2025) can be directly interpreted as the single state-dependent covariance propagation of a degenerate Kalman filter. In the case of nonlinear dynamics, enforcing the dependence of the error covariance $\mathbf{P}^{\mathrm{f}}$ on $\mathbf{x}^{\mathrm{f}}$ is even more beneficial since the covariance evolution now explicitly depends on $\mathbf{x}^{\mathrm{f}}$.

These theoretical results and proof experiments provide a rationale for the existence of a map $\mathbf{x}^{\mathrm{f}} \mapsto \mathbf{P}^{\mathrm{f}}(\mathbf{x}^{\mathrm{f}})$ and its importance in estimating the errors-of-the-day in DA schemes based on a single forecast state $N_{\mathrm{e}} = 1$.

(11) Lines 223 and 225: Clarify in what sense the following analysis is a "sensitivity analysis".

We now define the term explicitly. We use "sensitivity analysis" in the local differential sense: we inspect Jacobians of the learned analysis increment with respect to its two inputs, $\mathbf{x}$ and $\boldsymbol{\zeta}$, and use these derivatives to diagnose how a perturbation of the forecast state changes the locally inferred gain.

Changed text: ... To that end, we study the dependence of $\nabla_{\boldsymbol{\zeta}} \mathbf{a}_{\boldsymbol{\theta}}(\mathbf{x}, \boldsymbol{\zeta})_{|\boldsymbol{\zeta}=\mathbf{0}} \approx \mathbf{P}^{\mathrm{a}}(\mathbf{x})$ on $\mathbf{x}$, i.e. the forecast state. This can be seen as a linear-in-$\boldsymbol{\zeta}$ sensitivity analysis of $\mathbf{a}_{\boldsymbol{\theta}}$ with respect to its inputs.

(12) Line 228: The term "marginal gain" was confusing; consider clarifying or renaming it.

We like the term *marginal gain*. It draws from the economics vocabulary, defined as the incremental change of a variable when one of its dependencies is changed. Following your suggestion, we clarified the meaning. We have also double-checked the related terminology in the whole manuscript and made the required adjustments.

Changed text: Since $\boldsymbol{\Gamma}(\mathbf{x})$ is a proxy to $\nabla_{\mathbf{x}} \mathbf{P}^{\mathrm{a}}(\mathbf{x})$ as per Eq. (1), $\boldsymbol{\Gamma}(\mathbf{x})$ can be interpreted as the *marginal* variation of $\mathbf{P}^{\mathrm{a}}(\mathbf{x})$ when the forecast state $\mathbf{x}$ is perturbed. Under Gaussianity and linearity and assuming $\mathbf{H}^{\intercal}\mathbf{R}^{-1} = \mathbf{I}_{\mathrm{x}}$, $\mathbf{P}^{\mathrm{a}}(\mathbf{x})$ coincides with the Kalman gain. That is why we will call $\boldsymbol{\Gamma}(\mathbf{x})$ the *marginal gain tensor*, or simply *marginal gain*.

(13) Equation (10): Use ∂ rather than ∇ in the component-wise expression.

Isn't it what we did?

(14) Line 232: The identification with the Kalman gain needs linearity and optimally assumptions.

We agree and have added the required caveat.

Changed text: ... Under Gaussianity and linearity and assuming $\mathbf{H}^{\intercal}\mathbf{R}^{-1} = \mathbf{I}_{\mathrm{x}}$, $\mathbf{P}^{\mathrm{a}}(\mathbf{x})$ coincides with the Kalman gain. That is why we will call $\boldsymbol{\Gamma}(\mathbf{x})$ the *marginal gain tensor*, or simply *marginal gain*. ...

(15) Equation (12): Add the earlier and more DA-relevant references by Raanes et al. (2019) and Stordal et al. (2016).

Done. We now cite Raanes et al. (2019) and Stordal et al. (2016), but also Fillion et al. (2020), in this discussion and retain the ML reference only for its connection with perturbation-based explanations (LIME) and SmoothGrad.

(16) Line 250: Replace "second-order moment truncation" by "Gaussian".

Corrected.

(17) Equations (14)–(15): The notation was heavy for the approximation $\mathbb{E}_{\zeta}[\cdot] \simeq \cdot|_{\zeta=0}$, and the point did not rely on Stein's lemma.

This is true. However, what matters is that the intermediate steps of Eq. (15) suggest different routes to compute the mean marginal gain. For instance, $\mathbb{E}_{\mathbf{x}\sim\pi,\boldsymbol{\zeta}\sim\rho}[\boldsymbol{\Gamma}(\mathbf{x},\boldsymbol{\zeta})]$ suggests to compute a double Jacobian over $\mathbf{x}$ and $\boldsymbol{\zeta}$, whereas $\mathbb{E}_{\mathbf{x}\sim\pi}[\nabla_{\mathbf{x}}\mathbb{E}_{\boldsymbol{\zeta}\sim\rho}[\nabla_{\boldsymbol{\zeta}}\mathbf{a}_{\boldsymbol{\theta}}(\mathbf{x},\boldsymbol{\zeta})]]$ suggests to compute a single Jacobian over the outcome of a regression, the latter leveraging the Stein lemma. There is actually an interesting third route which uses an additional Stein lemma over the $\mathbf{x}$ variable, which we

successfully implemented and tested, but which may be too much for this paper. We have added a small paragraph to clarify and account for your remark.

Changed text: Note that Eq. (15) is a simple point estimator, which formalises that the averaged Jacobian $\mathbb{E}_{\boldsymbol{\zeta}}[\nabla_{\boldsymbol{\zeta}}\mathbf{a}_{\boldsymbol{\theta}}(\mathbf{x},\boldsymbol{\zeta})]$ is well approximated by $\nabla_{\boldsymbol{\zeta}}\mathbf{a}_{\boldsymbol{\theta}}(\mathbf{x},\boldsymbol{\zeta})|_{\boldsymbol{\zeta}=\mathbf{0}}$ when the distribution of projected innovations is concentrated near zero. The Stein lemma, which provides a regression-based estimator of this average, is not required per se to ascertain Eq. (15), but offers an alternative expression which we leverage in Appendix H.

(18) Line 270: The induced observation-space symmetry group should be isomorphic to G, making the assumption superfluous.

The reference to the appendix where the equivariance is proven pointed to the wrong appendix in the original manuscript. This has been corrected. We have removed the sentence where the induced symmetry is mentioned. Moreover, we have substantially revised the beginning of the appendix (now Appendix F), to clarify the concepts, at the risk of being a bit formal.

Changed text (in the introduction of the appendix): We wish to prove the equivariance of the marginal gain tensor, i.e. that for all $\mathbf{x}$ and $\boldsymbol{\zeta}$, $\boldsymbol{\Gamma}(g\circ\mathbf{x}, g\circ\boldsymbol{\zeta}) = g\otimes g^{\intercal}\otimes g^{\intercal}\circ\boldsymbol{\Gamma}(\mathbf{x},\boldsymbol{\zeta})$, under the action of an isometry $g\in\mathcal{G}$. The action of g on either state vector $\mathbf{x}$ or $\boldsymbol{\zeta}$ is represented by an orthogonal matrix $\mathbf{G}$: $g\circ\mathbf{x}=\mathbf{G}\mathbf{x}$, $g\circ\boldsymbol{\zeta}=\mathbf{G}\boldsymbol{\zeta}$.

We say that g preserves the fibres (preimages) of $\mathcal{H}_k$ if, for all $\mathbf{x}$, $\mathbf{x}'$ such that $\mathcal{H}_k(\mathbf{x})=\mathcal{H}_k(\mathbf{x}')$ one has $\mathcal{H}_k(g\circ\mathbf{x})=\mathcal{H}_k(g\circ\mathbf{x}')$. An isometry g_{y} defined over the observation space of $\mathcal{H}_k$ can then be associated to such g through $g_{\mathrm{y}}\circ\mathcal{H}_k(g^{\intercal}\circ(\cdot))=\mathcal{H}_k(\cdot)$. The action of such induced isometry g_{y} is represented by the orthogonal matrix $\mathbf{G}_{\mathrm{y}}$: $g_{\mathrm{y}}\circ\mathbf{y}=\mathbf{G}_{\mathrm{y}}\mathbf{y}$. Because $\mathbf{G}$ and $\mathbf{G}_{\mathrm{y}}$ are orthogonal, one has $\mathbf{G}^{-1}=\mathbf{G}^{\intercal}$ and $\mathbf{G}_{\mathrm{y}}^{-1}=\mathbf{G}_{\mathrm{y}}^{\intercal}$.

With these definitions in hand, we consider a maximal group of isometries $\mathcal{G}$ defined over the state space such that

- $\mathcal{G}$ preserves the fibres of all $\mathcal{H}_k$, inducing a group of isometries $\mathcal{G}_{\mathrm{y}}$ (possibly for each k). $\mathcal{G}_{\mathrm{y}}$ is isomorphic to $\mathcal{G}$,
- the associated $\mathcal{G}_{\mathrm{y}}$ satisfies $\mathbf{G}_{\mathrm{y}}\mathbf{R}_k\mathbf{G}_{\mathrm{y}}^{\intercal}=\mathbf{R}_k$, which accounts for any heteroscedasticity of the observation error statistics,
- $\mathcal{G}$ commutes with the autonomous dynamics, i.e. $\mathbf{G}\mathcal{M}(\mathbf{G}^{\intercal}(\cdot))=\mathcal{M}(\cdot)$.

A consequence of the equivariance with respect to $\mathcal{H}_k$, i.e. $\mathbf{G}_{\mathrm{y}}\mathcal{H}_k\left(\mathbf{G}^{\intercal}(\cdot)\right)=\mathcal{H}_k(\cdot)$, is $\mathbf{G}_{\mathrm{y}}\mathbf{H}_k\mathbf{G}^{\intercal}=\mathbf{H}_k$, readily obtained by differentiation.

(19) Section 3.2.2: Provide a less group-theoretical summary, perhaps explaining the construction as a circulant average.

We have followed your suggestion. The beginning of the section has been revised.

Changed text: An important way to reduce the computational cost and complexity of interpreting the mean marginal gain $\widetilde{\boldsymbol{\Gamma}}$ is to exploit the symmetries of the DA problem, when applicable. If the model variables are observed homogeneously and homoscedastically, then the probability laws defining the DA problem are invariant under these symmetries, and the corresponding analysis operator is equivariant. As a result, the components of $\widetilde{\boldsymbol{\Gamma}}$ related by symmetry are redundant, so that its interpretation can be restricted to a reduced set of representative components.

For instance, in the one-dimensional periodic L96 model with homogeneous and homoscedastic observations, a cyclic translation of the state variables induces the same cyclic translation of the observations, innovations, and analysis increments. The probability laws defining the DA problem are therefore invariant under the translation group, while the corresponding analysis operator is equivariant. Consequently, if the trained DAN preserves this symmetry, any trajectory-averaged sensitivity tensor must have a circulant structure: the information associated with different grid points is identical up to a cyclic shift. The group-theoretical argument below provides a formal justification for this circulant averaging.

(20) Equation (21): The notation made it hard to see why the result is a two-index object; the role of r was unclear.

We have changed the notation to make r explicitly fixed. The object $\overline{\mathbf{\Omega}}^{(r)}$ is now introduced as a matrix slice of the averaged 3-tensor with its third index fixed at the reference site r. We have also significantly reduced the usage of the bracket notation around tensors, throughout the manuscript.

Changed text: Leveraging the equivariance induced by the translational invariance of the L96 model, for a fixed reference site r, we define the *slice* matrix $\overline{\mathbf{\Omega}}^{(r)} \in \mathbb{R}^{N_x \times N_x}$ component-wise by

$$\overline{\Omega}^{(r)}_{ij} \triangleq \overline{\Gamma}_{ijr}. \tag{4}$$

The index r is not an additional free index of $\overline{\mathbf{\Omega}}^{(r)}$; it labels the chosen slice of the averaged 3-tensor.

(21) Line 343: The omission of the IEnKS was disappointing; existing benchmarks should at least be referenced.

The absence of the IEnKS is deliberate. How we explained this omission was actually sloppy and did not provide the more fundamental reasons why we chose to do so. This is now explained.

Changed text: The iterative ensemble Kalman smoother (Bocquet and Sakov, 2014; Raanes et al., 2019) has the potential to outperform those benchmarks, including for filtering metrics. However, it works over longer DA windows, as opposed to the DAN implemented here, so that such comparison would be biased. Actually, we have developed variants of DAN that work on extended DA windows similarly to the IEnKS, but reporting on them is beyond the goals of this paper.

(22) Line 366: Replace "passed" by "beyond".

Corrected.

(23) Figure 4 caption: Replace N by N_{iter}.

N_{iter} does appear when using the original pdf and using any pdf viewer we have access to. We have checked that all the required fonts are embedded in both the standalone figure and in the original manuscript pdf. Hence, the missing "iter" seems to be due to Copernicus' processing of the figures.

(24) Line 459: The claim that the DAN must learn the map $\mathbf{x}^{\mathrm{f}} \mapsto \mathbf{P}^{\mathrm{f}}(\mathbf{x}^{\mathrm{f}})$ was too strong in view of the poor performance of the linear-in-ζ DAN in nonlinear regimes.

We agree and have softened the conclusion. The linear-in-ζ DAN indicates that the map $x^{\mathrm{f}} \mapsto P^{\mathrm{f}}(x^{\mathrm{f}})$ is sufficient in the mild nonlinear regime, but Fig. 3 shows that it is not sufficient in stronger nonlinear regimes. The full DAN must therefore encode flow-dependent uncertainty information and nonlinear/non-Gaussian corrections, but it need not explicitly compute a covariance matrix.

Changed text: We have previously shown that, to achieve this level of performance, DAN must implicitly learn a map $\mathbf{x}^{\mathrm{f}} \mapsto \mathbf{P}^{\mathrm{a}}(\mathbf{x}^{\mathrm{f}})$, at least in the mildly nonlinear regime. Its existence is supported by a multiplicative ergodic theorem applied to the entire DA process viewed as a dynamical system.

(25) Unsorted: "Expansion" is confusing if the operation is a first-order regression; consider "linearisation".

We agree and have replaced "expansion" by "linearisation" wherever the intended meaning is a first-order local or regression-based approximation.

(26) Unsorted: The code should be available to the reviewers.

We preferred to submit this manuscript to *Nonlinear Processes in Geosciences* where the focus is on the methods, rather than to *Geoscientific Model Development* where the focus is on the implementation. However, upon acceptance, the core codes will be made public through a versioned repository with a persistent identifier.

(27) Unsorted: Add a dedicated comment on the performance of the linear-in-ζ DAN and the DAN with linear activation functions.

With linear activation functions only, DAN is essentially an affine map of its inputs and cannot build a rich state-dependent prior; its behaviour is therefore close to that of a static background method. The linear-in-$\boldsymbol{\zeta}$ DAN is more expressive because $\mathbf{P}^{\mathrm{a}}_{\theta}(\mathbf{x})$ depends nonlinearly on the forecast state, and it succeeds in the mild regime. Its degradation as Δ_t increases shows that a covariance-like dependence on $\mathbf{x}$ is not enough: the full DAN also uses nonlinear dependence on $\boldsymbol{\zeta}$ to represent non-Gaussian corrections.

A similar comment for the linear activation functions is actually already present in Sect. 4.1. And the subsection Sect. 4.3.1 is later dedicated to the performance and interpretation of the linear-in-$\boldsymbol{\zeta}$ DAN. It is not included in the discussion of the numerical results because it is not used as a benchmark but as a demonstration tool.

(28) Unsorted: Add a comment relating to moment closure in linear dynamics and why introducing covariance dependence on the state can be fruitful in nonlinear dynamics.

We believe you are referring to the discussion of Sect. 3.1.3, *On the importance of the* $\mathbf{x}^{\mathrm{f}} \mapsto \mathbf{P}^{\mathrm{f}}(\mathbf{x}^{\mathrm{f}})$ *map.* We agree with you that in the nonlinear dynamics case, enforcing the dependence of the error covariances on the state should be systematically helpful since the tangent linear model now depends on the state. However, even in the linear non-autonomous dynamics case, one expects the error covariance matrices to asymptotically collapse onto the unstable subspace which is time-dependent. This subspace can in theory be inferred from the state if the MET is applicable. Hence, nonlinearity is likely not the only reason why the dependence of the error covariances on the state is fruitful. Thank you for pointing it out and the suggestion. We have added a sentence in Sect. 3.1.3.

Changed text: In the case of nonlinear dynamics, enforcing the dependence of the error covariance $\mathbf{P}^{\mathrm{f}}$ on $\mathbf{x}^{\mathrm{f}}$ is even more beneficial since the covariance evolution now explicitly depends on $\mathbf{x}^{\mathrm{f}}$.

(29) Unsorted: Have the authors investigated benefits of operating DANs based on ensemble forecasts?

In Boc24, we tested ensemble variants of DANs and found no clear point-estimation accuracy gain in the L96 settings considered there once the single-state DAN was sufficiently trained. This solution was concomitant to a *feature collapse* of the neural network. We pointed out however, that a better solution with an explicit dependence on the ensemble was likely but would probably not yield a significantly better accuracy, since the single-state DAN is already on par with a well-tuned EnKF.

Differently, what we could test in the future would be based on a single-state trained DAN, to generate an analysis ensemble from $\mathbf{a}_{\boldsymbol{\theta}}$ (similarly to how we generate perturbations in the regression of $\mathbf{a}_{\boldsymbol{\theta}}$ to a Kalman gain matrix), and forecast it to generate an ensemble forecast. This might yield an accuracy improvement due to the ensemble forecasts, in place of single-state forecasts, when the dynamics are significantly nonlinear.

References

Bocquet, M. and Carrassi, A.: Four-dimensional ensemble variational data assimilation and the unstable subspace, Tellus A, 69, 1304 504, https://doi.org/10.1080/16000870.2017.1304504, 2017.

Bocquet, M. and Sakov, P.: An iterative ensemble Kalman smoother, Q. J. R. Meteorol. Soc., 140, 1521–1535, https://doi.org/10.1002/qj.2236, 2014.

Bocquet, M., Gurumoorthy, K. S., Apte, A., Carrassi, A., Grudzien, C., and Jones, C. K. R. T.: Degenerate Kalman filter error covariances and their convergence onto the unstable subspace, SIAM/ASA J. Uncertain. Quantif., 5, 304–333, https://doi.org/10.1137/16M1068712, 2017.

Bocquet, M., Farchi, A., Finn, T. S., Durand, C., Cheng, S., Chen, Y., Pasmans, I., and Carrassi, A.: Accurate deep learning-based filtering for chaotic dynamics by identifying instabilities without an ensemble, Chaos, 29, 091 104, https://doi.org/10.1063/5.0230837, 2024.

Carrassi, A., Bocquet, M., Demaeyer, J., Gruzien, C., Raanes, P. N., and Vannitsem, S.: Data assimilation for chaotic dynamics, in: Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. IV), edited by P., S. K. and X., L., pp. 1–42, Springer International Publishing, Cham, ISBN 978-3-030-77722-7, https://doi.org/10.1007/978-3-030-77722-7_1, 2022.

Daley, R.: Atmospheric Data Analysis, Cambridge University Press, New-York, ISBN 9780521458252, 1991.

Fillion, A., Bocquet, M., Gratton, S., Gürol, S., and Sakov, P.: An iterative ensemble Kalman smoother in presence of additive model error, SIAM/ASA J. Uncertain. Quantif., 8, 198–228, https://doi.org/10.1137/19M1244147, 2020.

Palatella, L., Carrassi, A., and Trevisan, A.: Lyapunov vectors and assimilation in the unstable subspace: theory and applications, J. Phys. A: Math. Theor., 46, 254 020, https://doi.org/10.1088/1751-8113/46/25/254020, 2013.

Raanes, P. N., Stordal, A. S., and Evensen, G.: Revising the stochastic iterative ensemble smoother, Nonlin. Processes Geophys., 26, 325–338, https://doi.org/10.5194/npg-26-325-2019, 2019.

Sakov, P.: On building the state error covariance from a state estimate, https://doi.org/10.48550/arXiv.2411.14809, 2025.

Stordal, A. S., Szklarz, S. P., and Leeuwenburgh, O.: A Theoretical look at Ensemble-Based Optimization in Reservoir Management, Mathematical Geosciences, 48, 399–417, https://doi.org/10.1007/s11004-015-9598-6, 2016.