Reviewer #1:

Minor Comments:

The authors have considered all of the remarks by the reviewers. The manuscript has been very much improved, in particular through the following actions:

1. Addition of many relevant references
2. Better description of aims and their differences with previous studies
3. Addition of a table describing experiments and assimilated observations
4. Incorporation of no observations data assimilation (NODA) experiments in the evaluation
5. Discussion focused on the well-defined aims.

I believe the manuscript can be published after the following remark is taken care of:
Figure 9: Why are there different number of points on the curves for different experiments? This artificially affects the AUC values.

We thank the reviewer for the positive feedback and for the constructive comments that have helped us improve the manuscript.

Regarding the comment on Figure 9, we apologize for the confusion. In fact, **all experiments use the same set of thresholds**, resulting in an identical number of ROC points across runs.. The apparent difference arises because, in some experiments, the ROC points cluster very closely near the origin (i.e., at low probability thresholds), causing them to overplot and appear as fewer distinct markers.

To address this and improve clarity, we have **added a note** in the figure caption explaining that all experiments use the same thresholds and that clustering of points reflects tightly grouped ROC values rather than differing point counts.

Now, the figure caption of Figure 9 reads as:

***"Figure 9**. ROC curves and AUC associated with the 3DVar (red and pink colors), EnKF (blue and cyan colors) and NODA (green color) for the 3-hour accumulated precipitation using (a) 1 mm and (b) 10 mm threshold and 6-hour accumulated precipitation using (c) 1 mm and (d) 10 mm threshold, computed over the entire inner domain. **Note:** all experiments employ the same set of probability thresholds; any apparent differences in the number of plotted points arise from clustering of ROC values at similar thresholds, not from differing data counts."*

We trust these changes resolve the concern.

Again, we want to appreciate the reviewer's feedback, as it has significantly improved the quality of the manuscript.

Reviewer #2:

Major Comments:

The literature has now been sufficient extended and includes more works on the topic. However, in lines 152–186, the authors state:

"However, at this stage, the system is already well-developed and likely impacting the population, limiting the effectiveness of DA in terms of forecast lead time. In such cases, the potential for early warnings and mitigation actions is significantly reduced, as there is little time left to respond and minimize socio-economic impacts. Despite its potential benefits, very few studies have explored the role of DA in the developing stage (e.g., Carrió et al., 2019; Carrió et al., 2022; Corrales et al., 2023), where assimilating observations before convection initiates could significantly improve forecast lead time, providing advanced warnings and allowing decision makers to act proactively."

I do not fully agree with this statement in its current form, as several studies in the literature have already addressed the improvement of forecasts through the assimilation prior to convection initiation, including in coastal Mediterranean areas. Otherwise, the practical value of such forecasts would be minimal or negligible.

We thank the reviewer for this comment and appreciate the opportunity to clarify this point. Our intent was to motivate this study, highlighting that most of the DA studies at convective-scales are focussed on the mature stage of convective weather events over well-observed land regions. In contrast, DA studies focussed on the development of convective systems over maritime areas, where observations are scarce, remain relatively scarce. One of the aims of this study is to fill this gap by comparing the performance of two widely used DA methods under these data-sparse, maritime conditions. To the best of our knowledge, only a handful of studies have explored analogous setups, combining extreme-case pre-convective assimilation, maritime observation constraints, and high-resolution modeling.

We have adjusted the manuscript to make this distinction clearer and to acknowledge existing work while emphasizing the novelty of our focus on pre‑convective DA in maritime extreme events.

"However, at this stage, the system is already well-developed and likely impacting the population, limiting the effectiveness of DA in terms of forecast lead time. In such cases, the potential for early warnings and mitigation actions is significantly reduced, as there is little time left to respond and minimize socio-economic impacts. Despite its potential benefits, only a handful of studies have explored the impact of DA using high-resolution numerical models in the developing stage (e.g., Carrió et al., 2019; Carrió et al., 2022; Corrales et al., 2023), and even fewer have done so over data-sparse maritime regions, where early assimilation could be most valuable, providing advanced warnings and allowing decision-makers to act proactively. This study fills that gap by directly comparing two widely used DA techniques – 3DVar and EnKF – in high-resolution, pre-convective assimilation experiments for two extreme weather events initiated over the sea affecting populated coastal regions in the Mediterranean basin."

Moreover, in point (b) among the stated objectives, the authors write:

"Investigate the potential of using 3DVar and EnKF in the developing phase, that is hours before the mature stage of convective systems are reached, to improve forecast lead time and warning capabilities for extreme weather events."

This may be a matter of how the work methodology is currently structured and presented, but as it stands, the manuscript does not clearly highlight a substantial difference from existing studies. I recommend specifying how many hours before the mature phase of the event the final assimilation cycle takes place. This would help demonstrate that assimilation is indeed performed well ahead of convection initiation. Previously, part of the validation was carried out only for the first hour immediately following the last assimilation cycle, which gave the impression that the forecast was not really anticipating the event by much. This may simply be clarified by better defining when the intense phase of the event begins with respect to the last assimilation cycle.

We thank the reviewer for this suggestion. To clarify the timing of our pre-convective assimilation, we have added the lead time relative to the onset of the mature convective phase. In the revised manuscript, Objective b) now reads as:

" b) Investigate the potential of using 3DVar and EnKF in the developing phase, specifically 12 hours before the mature stage of convective systems are reached, to improve forecast lead time and warning capabilities for extreme weather events."

Based on the simulations setup shown in Figure 6, assimilation continues for 24 hours before the start of the free forecast. Therefore, in the context of "improve forecast lead time and warning capabilities for extreme weather events", what matters is not the initialization time of the forecast itself, but the timing of the final assimilation cycle. In a "warning capabilities" perspective, the forecast would only be available after all observations have been acquired and assimilated, that is, after the last cycle.

We thank the reviewer for this clarification. To avoid confusion, we have removed "warning capabilities" and focused Objective (b) solely on forecast lead time. The revised objective now reads:

" b) Investigate the potential of using 3DVar and EnKF in the developing phase, specifically 12 hours before the mature stage of convective systems are reached, to improve forecast lead time."

In my view, this remains a crucial point to clarify or change before the publication can be accepted. If the last assimilation cycle is only a couple of hours ahead of the onset of the event, many studies have already investigated similar setups. In that case, I would suggest placing less emphasis on this specific aspect and focusing more directly on the comparison between 3DVar and EnKF, possibly removing this part from the discussion. Alternatively, rather than limiting the analysis to the forecast phase, the authors could consider including an investigation of the pre-convective stage during the assimilation cycles. This would help demonstrate how assimilation improves the model state and how cycling gradually corrects the model, thereby reducing the propagation of errors into the subsequent forecast.

We thank the reviewer for this insightful recommendation. In light of concerns about manuscript length and scope, we have removed the emphasis on the pre-convective assimilation window from our

discussion and refocused the objectives on the core comparison of 3DVar and EnKF forecast performance. The revised objective now reads as follows:

"On overall, this study aims at:

(a) Assessing the impact of 3DVar in comparison with the EnKF system to predict small-scale extreme weather events initiated over maritime regions with lack of in-situ observations.

(b) Compare the forecast impact from assimilating in-situ conventional observations in comparison to assimilating high spatial and temporal resolution data from remote sensing instruments.

(c) Provide a quantitative assessment between the different DA schemes by means of using several statistical verification methods."

We would also like to underscore the novelty of our work: to our knowledge, no previous studies have directly compared 3DVar and EnKF in high-resolution, convective-scale models applied to extreme events initiated over the sea, nor have they evaluated the specific combinations of Doppler radar and satellite observations that we assimilate. We believe this focused comparison offers a valuable contribution to the DA and NWP communities.

Minor comments:

1) NODA simulation: In Section 5 (Model set-up), the configuration of the NODA run is not mentioned. Is it the same as the one used with 3DVar?

Yes, NODA indeed uses the same model configuration as the 3DVar. Section 5 is not intended to introduce the different numerical experiments. Since Section 5 focuses on the general model setup rather than individual experiments, we added clarification in Section 6 where the experiments are introduced. The revised text in Section 6 now reads as follows:

"To quantify the benefits of assimilating different observation types with the 3DVar and EnKF DA schemes, a suite of numerical experiments is designed. First, a reference experiment without any data assimilation (NODA), using the same model configuration employed for the WRF experiments performed using 3DVar, is carried out at the regional scales considered in this study."

2) Lines 820-830: Regarding the computation of the background error covariance matrix, there is certainly no strict minimum duration, and if it is appropriately calculated, it can lead to improvements in the forecast. However, since the goal of this study is not only to demonstrate improvement compared to the NODA simulation but also to provide a comparison with a method such as EnKF, it becomes crucial that the covariance matrix is computed in the most robust way possible and based on a sufficiently large statistical sample — ideally at least one month, if not a full season (in operational settings, multiple years are often used to ensure robustness). This is particularly important as the background error covariance is one of the cornerstones of variational assimilation (Stanesic et al., 2019). The fact that it was computed over such a short period remains a limitation of the 3DVar approach and should be acknowledged in the comparison with EnKF in the text.

Stanesic A, Horvath K, Keresturi E. Comparison of NMC and Ensemble-Based Climatological Background-Error Covariances in an Operational Limited-Area Data Assimilation System. Atmosphere. 2019; 10(10):570. https://doi.org/10.3390/atmos10100570

We thank the reviewer for highlighting the importance of a robust background error covariance (BEC) matrix, and agree that longer accumulation periods (e.g., a month or season) generally contribute to more reliable variational analysis. However, the choice in this study to use a BEC matrix computed on a short period is based on the CETEMPS extensive experience and previous results (e.g., Hung et al. 2023; Fitzpatrick et al., 2007; Mazzarella et al., 2020, 2021). Moreover, a further improvement in the calculation of the BEC is obtained by applying the recently developed enhancement, which is the inclusion of the CV7 option in WRFDA. This option is particularly beneficial for improving precipitation estimates and the assimilation of radar reflectivity data (Wang et al., 2013; Li et al., 2016; Shen et al., 2022; Ferrer Hernandez et al., 2022), as CV7 utilizes orthogonal functions instead of a vertical recursive filter, leading to a more accurate representation of error correlations and potentially improving the quality of our BEC despite the shorter temporal sampling. We would like to emphasize that at CETEMPS the DA for assimilating both radar data and conventional data is based on a BEC matrix computed operationally on the short term because showed the best results on the forecasts.

We have included this discussion in the revised text and now reads as follows:

"*... In this study, we build the 3DVar **B** matrix over a two-week period, in line with our operational experience running 3DVar and previous demonstrations of its benefits (Hung et al. 2023; Fitzpatrick et al., 2007; Mazzarella et al., 2020, 2021). To enhance **B**'s quality despite this relatively short sampling window, we activate the CV7 option in WRFDA. This option uses empirical orthogonal functions (EOFs) to represent vertical covariances instead of the traditional recursive filter, which has proven particularly beneficial for radar-reflectivity assimilation and subsequent precipitation forecast improvements (Wang et al., 2013; Li et al., 2016; Shen et al., 2022; Ferrer Hernandez et al., 2022). In our configuration, the CV7 control variables (i.e., u, v, temperature, pseudo-relative humidity and surface pressure), are defined in EOF space, ensuring a compact yet accurate representation of error structures. We use the CV7 option to generate the **B** matrix for both case studies. In addition, the weak penalty constraint (WPEC) option (Li et al., 2015) in WRFDA has also been activated to further improve the balance between the wind and thermodynamic state variables, enforcing the quasi-gradient balance on the analysis field."*

3) Lines 1132-1141: R3 is located near the edges of the domain used and relies solely on in-situ sensors for assimilation corrections, as it lies outside the radar coverage area. This could be another reason for the result obtained.

We thank the reviewer for this observation. In the revised text, we have updated the discussion of R3 to include this additional factor:

"In R3, the results show an unexpected behavior when using the moderate threshold (5 mm·h-1) (Fig. 8c), where NODA outperforms DA simulations during the first few hours. This anomaly could be attributed to three factors: (1) the use of a moderate precipitation threshold, which may not capture significant precipitation differences; (2) minimal precipitation in R3 during the initial forecast hours, since the deep convection system had not yet reached this region; and (3) the location of R3 near the domain edges, where it relies solely on in-situ observations for assimilation corrections, as it falls outside the radar coverage area."

Reviewer #3:

Minor Comments:

I am pleased with the revisions, which fully addressed the points raised and makes the outcomes of the study clear. Now the manuscript is in good shape and I recommend minor textual revisions, with no need for another review round.

We thank the reviewer for all the comments and suggestions that helped us to significantly improve the quality of this study.

It could be noted that some predictability was already present in the noDA experiment since ECMWF performed assimilation at a somewhat larger spatial scale?

We thank the reviewer for this insightful observation. Indeed, the **NODA** experiment benefits from the **large-scale ECMWF analysis** on which all ensemble configurations are initialized. Because **all simulations (NODA, 3DVar, and EnKF)** share the same ECMWF background, any advantage conferred by the ECMWF's large-scale assimilation is **common to all experiments**. Highlighting this point in detail risks diverting focus from our primary comparison of the DA methodologies themselves. Consequently, we have chosen not to expand on this aspect in the manuscript but have instead clarified in the text that all runs use the same ECMWF initial conditions.

The beginning of **Section 6** now reads as:

"To quantify the benefits of assimilating different observation types with the 3DVar and EnKF DA schemes, a suite of numerical experiments is designed. First, a reference experiment without any data assimilation (NODA), using the same model configuration employed for 3DVar, is performed at the regional scales considered in this study. Building on this, several numerical experiments, each differing only in the type of observations assimilated to isolate and compare their impacts on forecast skill, are performed."

L143: Wording: "solve" should probably mean "resolve"?

Done.

L148: Rephrase "In this context, which DA method is more suitable?". For example: "Given limited computational resources, it is unclear which DA method is more accurate." However, as you know, the answer might depend on how big the resources are.

Done. We have rephrased the previous sentence as follows:

"Determining which DA method yields greater accuracy – 3DVar using an ad hoc background error covariance matrix versus EnKF with a flow-dependent low-rank background error covariance derived from a finite ensemble – remains challenging under constrained computational resources."

L427-428: "The assimilation of each observation results in a reduction of the ensemble spread, attributed to using a reduced-moderate ensemble size" Confusing. Assimilating an observation reduces the analysis variance in variational and Kalman filter assimilation methods. If you want to motivate adaptive inflation, you could say that the small/finite ensemble sizes shrink the ensemble spread more than it should.

Agree. We have changed it to the following:

"Assimilating observations inherently reduces analysis variance in both variational and Kalman filter frameworks. Small ensemble sizes tend to overly collapse the ensemble spread (Anderson and Anderson, 1999). To mitigate this underdispersion and maintain realistic ensemble variance, a spatially varying adaptive inflation technique (Anderson and Collins, 2007; Anderson et al., 2009) is applied to the prior ensemble before assimilating the observations."

L429ff: Which technique was applied: spatially varying or homogeneous? According to the citation it was spatially varying?

Right. The spatially varying technique was applied. We have added this clarification in the manuscript:

"To mitigate this underdispersion and maintain realistic ensemble variance, a spatially varying adaptive inflation technique (Anderson and Collins, 2007; Anderson et al., 2009) is applied to the prior ensemble before assimilating the observations."

Fig 8(i): Data for 3DVar CNTRL is missing. Lower panels: RMSE unit is mm/h?

We thank the reviewer for noticing this missing. We have added the missing curve for 3DVar CNTRL and added the RMSE units.

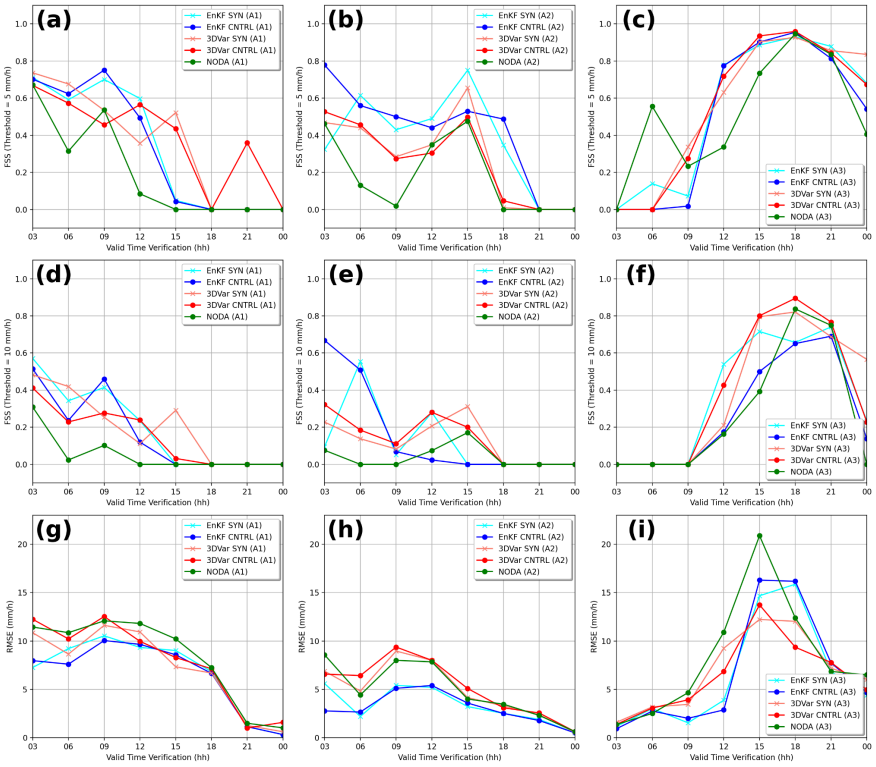

Fig 13: Caption: "Probability of cyclone center occurrence" add for example "(within 20 km)" Check if the values on the colorbars are correct: are all values below 1% or is 0.16 actually 16%? Tick values should appear once on the colorbar. I guess there is a rounding in place.

We thank the reviewer for these careful checks.

1. **Caption:** We have updated the figure caption to read **"Probability of cyclone center occurrence (within 20 km)"**.

2. **Colorbar values:** The tick labels on the colorbars have been corrected to display each unique value only once after removing unintended rounding.

These changes ensure the figure accurately conveys the intended information.

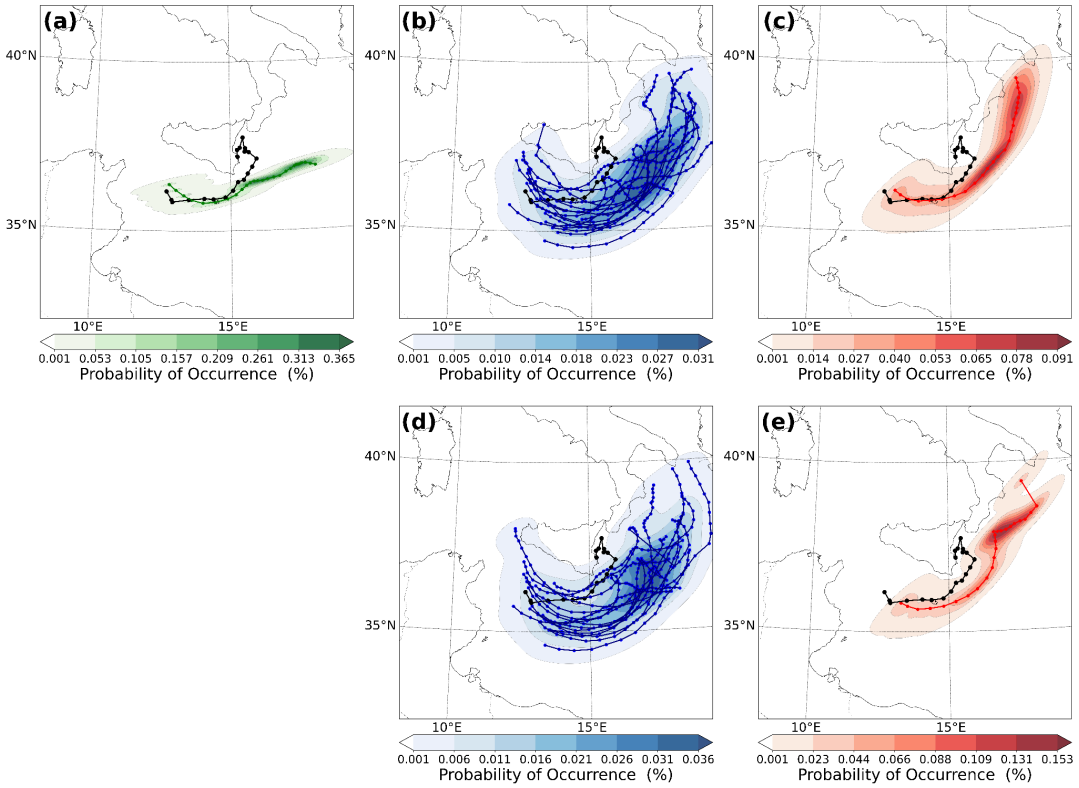

Figure 13. *Probability of cyclone center occurrence (within 20 km) computed using Gaussian KDE for (a) NODA, (b) EnKF (SYN), (c) 3DVar (SYN), (d) EnKF (CNTRL) and (e) 3DVar (CNTRL), from 11 UTC 7 November to 12 UTC 8 November 2014. Qendresa's trajectory observed via satellite imagery is depicted in black.*

L980-981: Ensemble members are deterministic forecasts, right? If so, replace "deterministic numerical weather models" by "the NODA forecast".

Done.

L1060: "much less" ... can you quantify that approximately?

We thank the reviewer for this request for clarification. Rather than leave the comparison qualitative, we have replaced "much less" with an approximate estimate based on the **ensemble size used in our EnKF experiments**. The main text has modified as follows:

"Although the EnKF technique has shown in general better performance against the 3DVar for the two extreme weather events analyzed in this study, it is also important to account for the computational resources required by each method. The EnKF requires approximately 36 times more model integrations per cycle than 3DVar's single forecast, in addition to the overhead of computing ensemble updates. This makes the 3DVar appealing because it is much faster and cheaper than the EnKF, and it makes this technique particularly suitable for operational purposes at the small weather forecast centers."

L1062-1034: "it does not need either to simulate model trajectories between the assimilation of a set of observations at time t1 and the subsequent set of observations valid at t2" Confusing. Do you mean that 3D-Var simulates one model trajectory, while the EnKF needed to simulate 36 trajectories?

We thank the reviewer for pointing out the confusion. However, in our previous answer we already decided to remove the entire sentence, so this issue no longer arises in the new version of the manuscript.

L900-903: "very different" feels vague. Better to be specific. "some members could completely fail in the prediction of the weather event": Does it mean that some members did not predict the existence of a cyclone but just unorganized convection?

We thank the reviewer for this suggestion. To provide greater clarity, we have replaced the vague phrasing with specific descriptions of member behavior. In the revised manuscript, the text now reads:

"The low predictability of Qendresa and the high sensitivity to physical parameterizations produce substantial spread in ensemble behavior: some members capture the cyclone's closed circulation and track reasonably well, while others fail to develop a coherent low-pressure core, instead producing only disorganized or weak convective cells. Consequently, these poorly performing members may entirely miss the medicane's formation or misplace its center, leading to large errors in both track and intensity forecasts."

L903-904: " In this situation, our small-to-moderate ensemble will probably produce a poor flowdependent background error covariance matrix": If it is somewhat probable that there is no cyclone, then this information should be in the background error covariance, I would say. However, you might mean that large uncertainty/spread leads to substantial nonlinearity, which is detrimental to the analysis accuracy of the EnKF.

We thank the reviewer for this insightful comment. You are correct that, in principle, a perfectly represented background error covariance matrix would reflect all uncertainties, including the possibility of no cyclone. In practice, however, finite-sized ensembles suffer from sampling error, which leads to spurious correlations and underdispersion, issues that are exacerbated when the underlying model forecasts fail to represent the event accurately.

To clarify, we have revised the text to read:

"In this situation, our small-to-moderate ensemble size exacerbates sampling error, yielding spurious background error covariances that degrade analysis accuracy in the EnKF. These errors become particularly problematic when the numerical model mispredicts the event, since the ensemble members no longer provide a reliable representation of flow-dependent uncertainty."

L905-907: " for which the ensemble mean will be smoothed out significantly". The ensemble mean should not be expected to be a good forecast of the true state, in case the distribution is non-Gaussian, which it will be for extreme precipitation and the cyclone's pressure field.

Agree. Since this sentence has been removed in the revised manuscript, we confirm that no further action is needed on this point.

L909-911: it is important to note that although the ensemble mean of the EnKF_SYN is not correctly reproducing the intensification of Qendresa, some of the ensemble members very well reproduce the

observed MSLP both in deepening and timing" In the light of the above comment, it should not be unexpected that averaging removes extreme values.

Agree.

L1069: "the 3DVar performs better than the EnKF ensemble mean" Yes, but then again, this is likely for extreme events because the ensemble mean is averaging over the skewed probability density function. I suggest rephrasing, since obviously in probabilistic metrics, like ROC/AUC, the situation is reversed.

We appreciate the reviewer for pointing out this. We have modified the main text as follows:

"An interesting result of this study is that, for highly non-Gaussian extreme events the deterministic 3DVar forecast can occasionally outperform the EnKF ensemble mean in terms of point forecasts (e.g., minimum central pressure), because averaging across ensemble members tends to smooth out the tails of a skewed probability distribution. In contrast, probabilistic metrics like ROC/AUC consistently favor the EnKF, reflecting its superior ability to capture forecast uncertainty. We attribute these contrasting behaviors to the different approaches to background error covariances: 3DVar employs a static covariance, while EnKF uses a flow-dependent covariance estimated from a finite ensemble. To combine the strengths of both methods, a hybrid error covariance approach—where the forecast error covariance matrix is formed by linearly blending the EnKF's ensemble-derived covariances with the 3DVar's static climatological covariances—may offer improved forecast skill for convective-scale extreme events."