Response to the reviewers of the paper "The operational 3DEnVar data assimilation scheme for the Meteo-France convective scale AROME-France"

Pierre Brousseau, Valerie Vogt, Etienne Arbogast, Maud Martet, Guillaume Thomas, Loïk Berre

February 2026

The authors would like to thank Christoph Schraff for the careful reading of the manuscript and his very constructive comments improving the manuscript quality. A point-by-point answer follows and the manuscript has been modified accordingly (red text in the track-changes file).

Please, note that in order to answer to the reviewer 2 comments :

- a new figure named figure 1 will be added in the manuscript and thus, the numbers of all the following figures are shifted.
- the number and the order of the subsections in section 2 have been modified.
- In section 4 and 5, the 3dv, 3dev and 3devi experiments will be respectively renamed 3DVar, EnVar and EnVarI experiments.

1 General comments

The Manuscript presents the work carried out to define the operational configuration of a new data assimilation scheme for the convective-scale model AROME at Meteo-France. This includes a number of sensitivity tests as well as the description of the operational configuration. It is shown that the new scheme based on 3DEnVar, using flow-dependent ensemble-based instead of climatological background error covariances, outperforms the old operational 3DVar scheme both in terms of statistical scores and for different kinds of individual severe weather events. While the manuscript does not directly address a better understanding of natural hazards, it does provide a significant contribution for an improved operational prediction of such hazards. 3DEnVar is a state-of-the-art method also used at other centres for both global and regional data assimilation, and the sensitivity experiments provide relevant information and valuable insight in its application at the convective scale. Generally, the work has good scientific originality, high relevance and quality, and the manuscript is well-written, well-referenced, and well-structured.

2 Minor specific comments

1. **Line 203: "... all sensitivity tests will be compared using the diagnostics that have been found to be representative of the obtained results." These diagnostics only include results from the 1-hourly DA cycle. In the end however, the purpose of the DA in the context of operational NWP is to provide initial conditions which result in the best forecast quality (in the case of AROME-France with lead times larger than 3 h since shorter lead times are available in real time only after their validity time). I think, the decision to use only diagnostics from the DA cycle for all the sensitivity experiments (except for lines 297 – 298) deserves more explanation here.**

   Indeed, the diagnostics used in this section are only informative on the behavior of the data assimilation cycle itself and on the quality of the 1 h range forecast used as the background. However, the impacts of the changes in these sensitivity experiments have also been evaluated on longer forecast ranges (6, 12 and 18 h). The results of these investigations are consistent with the results discussed in section 3 for the 1 h forecast and we have decided not to show them, in order to clarify this section which is already quite complicated. This explanation has been added at the end of section 3.1 :

   "Scores of 1 hr accumulated precipitation simulated by the 1 hr range forecasts of the DA cycle, as well as scores of longer accumulated precipitation (6 hr, 12 hr and 18 hr) have also been calculated. The results of these investigations on the sensitivity experiments are consistent with the results discussed below and

will not be shown. The use of 24 hr accumulated precipitation also allows a wide range of accumulated precipitation thresholds to be sampled enough to provide statistically significant results."

2. **Lines 211 – 213: I agree, simulated precipitation, being the result of many processes, can be a good indicator of the general behaviour of the system. However, (sums of) 1-h precipitation accumulation in the 1-h DA cycle may be strongly affected by (short-lived) spin-up effects. These may or may not have a significant impact on the forecast quality at later lead times. (In Section 3.3, e.g., the spin-up seen in surface pressure tendencies appears to have only a limited impact on the forecast quality, at least in the sense that experiments 3dev and 3devi perform fairly similar.) From that point of view, I would think that these 24-h precipitation sums of the DA cycle can be a valuable, but possibly quite incomplete indicator of the system behaviour with respect to precipitation. This should also be addressed (perhaps simply by further explanations, e.g. that the authors also looked at precipitation at longer lead times, but found similar results in the sensitivity tests (i.e. when comparing experiments) as in the diagnostic shown).**

   As also explained to reviewer 2 in point 2.15, we have also investigated the quality of 1 h accumulated precipitations simulated by 1 h forecasts of the DA cycle (without summation) and found very consistent results. We chose to show 24 h summation of 1 h forecasts because it allows a wide range of accumulated precipitation thresholds and then precipitation intensity (from 0.5 to 10 mm in 24 h) to be sampled enough to provide statistically significant results. As explained above, we have also investigated longer forecast ranges (6 h, 12 h and 18 h) of longer forecasts initialized at 00 and 12 UTC for all the sensitivity studies and also found results which are very consistent with those shown in section 3.

   We add some information about this in the text (see above).

3. **Lines 200 – 201: "... the localization length scale appears to be the most important parameter to investigate first." In my view, it need not be the most important parameter to be investigated and tuned first. It could also be the parameter for which the decision is most clear.**

   We agree with this comment and we propose to reformulate the sentence as follow :

   " Three different horizontal length scales for the horizontal localization have been first evaluated. Then, different experiments have been performed to evaluate the sensitivity to the vertical localization, inflation, ... "

   I am thinking here of the application of IAU which is shown to greatly reduce imbalances leading to spin-down of surface pressure tendency. Even though the impact of this on the forecast scores is shown to be quite limited (mostly slightly positive, but mostly not statistically significant), I would see this reduction of imbalances as a benefit on its own, which in the end proved to avoid the rare crashes of model simulations occurring without IAU. Hence, in the end, the chosen operational configuration includes IAU. This point may be relevant in the context of lines 268 – 273, e.g.: "The better results obtained in this study with shorter localization length scales than the optimal ones diagnosed with Ménétrier et al. (2015) could be explained by the fact that, while filtering sampled covariances, the localization helps to filter the numerical noise due to imbalances in the analysis and thus improves performances of the hourly DA cycle used. ...". These mentioned results in Section 3 were obtained without IAU, and my question is, whether this is also true to the same extent with IAU? In other words, would the optimal length scales be the same or larger when IAU is applied? Did the authors perform also sensitivity tests on the optimal localisation length scales with IAU?

   Yes, we also have performed localization length sensitivity experiments using IAU and found the same optimal localization length scale. We have added information in the article at the end of section 3.4 :

   "Finally, different combinations of values and mixture of the previous components have also been evaluated, but associated impacts are not detailed here. For example, horizontal localization sensitivity experiments using IAU have also been performed, leading to the same conclusion concerning the optimal lengths. Localization, by filtering sampled covariances, helps to improve analysis performances and, thus, performances of subsequent forecasts, while it has a weak impact on the spin-up reduction diagnosed on surface pressure tendencies. IAU filters numerical noise introduced by imbalanced analysis increments into the model integration during the forecast step : it has a strong impact on the surface pressure tendency but a weaker one on system performances. Localization and IAU do not filter the same things ; they have different impacts on forecast performances and on model stability, while both are useful for the system robustness."

4. **Lines 260 – 262: "Spin-up can also be referred as overestimated precipitation in the first few hours .... This aspect has already been investigated in section 3.2 and illustrated in**

figure 3...” This is not strictly true, because e.g. figure 3 shows HSS which is a skill score and not a measure of the bias which would be needed as an indicator of spin-up (if it can be referred as overestimated precipitation). Furthermore, it could be clarified that (a) ‘overestimated’ should be understood here in comparison with longer lead times, not in comparison to observations, and (b) underestimation (in this sense) could also be a result of spin-up effects.

We agree with this comment and we have chosen to remove these sentences from this section. The comment on the fact that the increase of the spin-up in the 3DEnVar experiments shown on this diagnostic is not detrimental to the general behavior of the model and its performances has been moved to section 4, which evaluates the selected configuration over a long time period.

3 Technical comments:

1. **Line 24: “using the Berre (2000) formulation”**

   Done

2. **Lines 27 – 28: FGAT: did this imply that observations were used at higher resolution. Any references?**

   Yes, the performed experiments used a 3h DA cycle. Observations from static platforms (SEVIRI radiances, screen-level observations, ...) were assimilated at a higher temporal resolution : every hour in the FGAT experiment instead of every 3h in the reference. FGAT allows for a more consistent estimation of O-B for observations onboard moving platform. It also allows to estimate different O-B's for a same observation on a static platform along the assimilation window, but the different innovations are considered at the same time (the center of the assimilation window) leading to a kind of temporal average on the associated analysis increment and smoothing the temporal information along the assimilation window. The performances of such experiments were degraded (Brousseau 2012 [1]).

   The text has been modified :

   ”the lack of temporal dimension within the assimilation window in the 3D scheme, which restricts the frequency of assimilated observations to the assimilation cycle frequency, as well as the temporal resolution of the comparison of the observation to the background. If the use of the First Guess at Appropriate Time (FGAT) approach may address the latter issue, it does not yield any improvements regarding the former (Brousseau 2012).

3. **Line 38: “allow the numerical cost”**

   The sentence has been replaced by *”keep the numerical cost within acceptable limits”.*

4. **Line 55: replace “counterpart” by e.g. “downside” or “drawback”**

   Done

5. **Line 73: “A hybrid”**

   Done

6. **Line 93: possibly better to replace “general scores” by “statistical scores”**

   Done

7. **Line 94: I do not understand what is meant by “original” approach**

   This point has been also raised by reviewer 2 and corrected. ”Original” was employed in the text to refer to the lagged-ensemble approach used to increase the sample size of the evaluated forecasts. For each meteorological issue (winter storm, fog and HPEs), different events are investigated, and for each event, the evaluation of the 3DEnVar scheme is done using numerous forecasts (at least 16) initialized at different analysis times in order to increase the investigation robustness.

   The wording ” an original approach” is misleading and has been removed in order to avoid to add more information in the introduction.

8. **Line 97: “numerical weather prediction system”**

   Done

9. **Line 112: This implies that PEARO runs at 1.3 km, while the EDA is at 3.2 km. It is worth mentioning how exactly the initial conditions of PEARO are obtained. Important to know also for lines 555 – 556.**

   The 24 perturbed initial atmospheric conditions used by AROME-EPS are generated as follows :

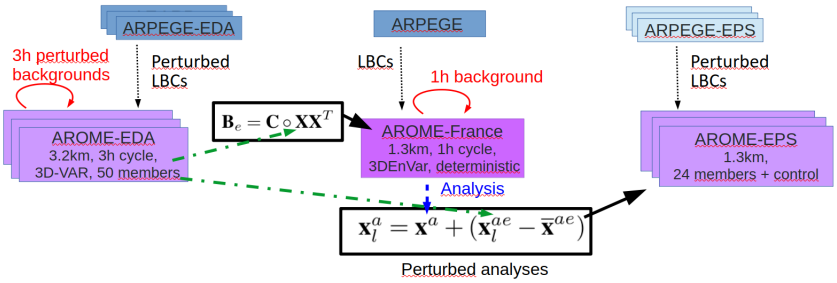


Figure 1: Schematic representation of the data exchanges between the main components of the Météo-France NWP operational system (deterministic, EDA and EPS versions) for the global model ARPEGE (blue rectangles) and the LAM AROME-France (purple rectangles). $\mathbf{B}_e$ is the ensemble background error covariance matrix, $\mathbf{C}$ the localization matrix, " $\circ$ " the Schur product and $\mathbf{X}$ the perturbation matrix provided by the AROME-EDA. $\mathbf{x}$ is a model state, $\overline{\mathbf{x}}$ the ensemble mean. Subscript l stands for an ensemble member and superscript a (resp. ae) for analysis (resp. AROME-EDA model state).

- 24 random draws of $(\mathbf{x}_l^{ae} - \overline{\mathbf{x}}^{ae})$ are selected from the 50 available differences (see new figure 1). Here, the superscript ae denotes an AROME-EDA state interpolated onto the 1.3 km horizontal grid, the subscript l indexes an AROME-EDA member, and $\overline{\mathbf{x}}^{ae}$ represents the ensemble mean.
- Each of the selected differences is then added to the AROME-France deterministic analysis, which is performed at 1.3 km resolution.

This information has been explicitly included in the new section 2.2 "Météo-France operational numerical weather prediction system and data flow" (previously section 2.1)

"For AROME-EPS, 24 perturbed initial atmospheric conditions are obtained by adding, to the AROME-France deterministic analysis ($\boldsymbol{x}^a$), 24 random draws among the 50 available differences ($\boldsymbol{x}_l^{ae} - \overline{\boldsymbol{x}}^{ae}$) from AROME-EDA."

10. **Line 127 – 128: for a better understanding: "These long forecasts (denoted 'production' in figure 1) have . . . "**

"production" has been replaced by "long forecasts" in Fig. 2.

11. **Line 131: I would think: "the IAU is not only used for its filtering purposes, but it is also employed to update the forecast with more recent observed information." More importantly, it should be clarified in Section 3 and 4 that H+1 observations and H+1 analysis increments (line 129) are not used in the sensitivity experiments with IAU (experiments "IAU" respc. "3devi") in order to allow for a fair comparison with the other experiments (or are they used?). I.e. that the same set of observations are input to all the sensitivity experiments.**

Indeed, section 3 only uses 1h forecasts of the hourly DA cycle and IAU is not used in the sensitivity experiments except in the IAU one. The comparison between experiments is fair, since in each experiment H+1 observations are never used in the 1h forecast of the DA cycle.

In section 4, it is only in the EnVarI (previously 3devi) experiment that the IAU is used in the 1h forecast of the DA cycle. In EnVarI (previously 3devi), IAU is used twice in the long forecasts, first to initialise the forecast with the H analysis, and also to update it with the H+1 analysis. In 3DVar (previously 3dv) and EnVar (previously 3dev) experiments, IAU is used only to update the long forecasts with the H+1 analysis. Here again, the comparison is fair : the long forecasts are updated with H+1 analysis increment in all experiments.

This is now stated in the article :

- at the end of the description of the sensitivity experiments at the end of the section 3.1 :
  "As this section relies on the investigation of forecasts from the DA cycle, IAU is only used in the corresponding experiment to initialize the forecast with the analysis performed at H. None experiment uses IAU to update the forecast initialized at H with the H+1 analysis as explained in section 2.2 for long forecasts"
- at the beginning of section 4 :
  "This section presents the results of the comparison between the 3D-Var (3DVar) and 3DEnVar experiments (EnVar and EnVarI) retained after the previous section over a 5.5 months period using different scores. In this section, the long forecasts initialized at H are also updated with the analysis performed at H+1 using IAU in the three experiments. Only the EnVarI experiment uses IAU to initialize the forecast with the H analysis"

12. **Line 141: replace "H+3" by "H-3" (?). H+3 would imply that the 9 UTC AROME-EDA uses 0 UTC AEARP lateral BC which would contradict figure 1.**

    Done

13. **Line 153: "... and while different uncompleted ..."**

    Done

14. **Figure 1: Elsewhere, the AROME EDA is named PEARO instead of AEARO.**

    Actually, PEARO is the AROME-EPS and AEARO the AROME-EDA. The use of ARPEGE, ARPEGE-EPS, ARPEGE-EDA, AROME-France, AROME-EPS and AROME-EDA has been generalized is the article (text and figure) and PEARO and AEARO have been removed in order to avoid confusion.

15. **Figure 1: For clarity, please explain "guess" somewhere.**

    "Guess" has been replaced by "background" as suggested by reviewer 2.

16. **Line 155: " ... ARPEGE DA system based on the OOPS code with all the ingredients ..." (?)**

    The sentence has been reworded as follows :

    "While previous versions of the code included prototypes for initial investigations and developments, the cy48 version is fully operational and enables the execution of AROME and ARPEGE DA systems with all the features present in the original FORTRAN code, including all operational observation types, variational bias correction and digital filters."

17. **Line 185: "control variable" (or rather: control vector?): should be explained what it is.**

    Done

18. **Line 185: "..- two horizontal wind components..."**

    Done

19. **Figure 2 caption: "... indicate larger (resp. smaller) RMS values ..."**

    Done

20. **Lin 235: "... show very similar performances ..."**

    Done

21. **Line 281: T, q, Ps have not yet been defined. Could be defined e.g. on line 186.**

    Done

22. **Figures 5 caption: remove "specific or"; in Figures 5 and 7, I would replace "Bold numbers" simply by "Numbers", because in Figures 2, 3 one has to distinguish between normal (not bold) numbers and bold numbers, while in Figures 5, 6, 7 when it says "bold numbers" I am trying to distinguish between bold and normal numbers but I cannot find any normal numbers.**

    Done

23. **Figure 5 and 6: I do not understand how a difference of 0.0 can be a statistically significant difference (by bootstrap). Please explain how this can happen, or what it means.**

    Actually, in order to make these figures more legible, RMS difference values are rounded to the nearest tenth. 0.0 values in bold number come from small differences, for exemple 0.02, which are statistically significant but rounded at 0.0 in the plotting script. These 0.0 values have been removed from the figures.

24. **Line 353 / Figure 6: "...IAU ... improves the analysis fit to observations." Do you have any idea why this is? Does it have to do with spin-up effects? What about the RMS Diff to the first guess (1-h forecast) which we cannot see in figure 6?**

    Yes, the improvement of the analysis fit to the observations in the EnVarI (previously 3devi) experiment is due to the reduction of the RMS (O-B) thanks to the use of IAU and the associated reduction of the spin-up. This information has been added in the text.

    This improvement is not obvious in old Fig. 2 (new Fig. 3) but appears stronger in old Fig. 6 (new Fig. 7) due to different verifying observations and a different sample (period, hour of the day). See next comment for more details.

25. **Figure 6, figure 2: The application of IAU is shown to lead to clearly smaller O-A against radiosondes in Figure 6, but hardly smaller O-A against aircraft data in Figure 2. What is the reason for this difference? Is it the different verifying observations? Or the different period? Or are there relevant discrepancies between (a) the differences between experiments 3devi and 3dev (in Figure 6) on the one hand and (b) the differences between experiments IAU and Loc h25-150 on the other hand? Please address this also in the manuscript.**

    There is no relevant discrepancy between (a) the differences between experiments EnVarI (previously 3devi) and EnVar (previously 3dev) used in old Fig. 6 (new Fig. 7) on the one hand and (b) the differences between experiments IAU and Loc h25-150 on the other hand. In both cases, the only difference between the experiment and the reference is the use of IAU to initialize the new forecast with the new analysis. The different impacts on the RMS(O-A) values due to the use of IAU, as observed in old Fig. 2 and 6 (new Fig. 3 and 7), can be attributed to differences in both the verifying observations and the sample characteristics (period, hour of the day) :

    - old Fig. 6 (new Fig. 7) shows relative differences of RMS(O-A) against radio-soundings valid at 00 UTC, averaged from 15 of September 2020 to the 1st of March 2021 ;
    - old Fig. 2 (new Fig. 3) shows relative differences of RMS(O-A) against aircraft measurements, averaged over all hourly analysis from 15 of September 2020 to 15 of October 2020.

    Relative differences of RMS(O-A) against aircraft measurements only valid at 00 UTC from the 15 of septembre 2020 to the 1st of March 2021 indicate values more consistent with those of old Fig. 6, reaching -3.5% for temperature at 500hPa (to be compared to -1.0 in old Fig. 2 and -4.7 in old Fig. 6). Moreover, as MODE-S observations are not used in this study, the number of aircraft measurements at 00 UTC, even if it is accumulated over a 5 month period, is very weak over the AROME-France domain (around only 300 observations at 500 hPa over the whole period) and the results are not statistically significant.

26. **Line 358: Maybe it is worth to briefly explain "double penalty". Something like: " ... which affects classical local scores if e.g. small-scale phenomena like convective cells are predicted well but with some dislocation." (Or provide a reference.)**

    Done . The reference Amodei et al. 2015 [2] explains the "double penalty" and how the BSS deals with this issue.

27. **Line 359: replace "excedence" by "exceedance"**

    Done

28. **Line 363: For clarity, I would write: "This neighbourhood is used to compute the AROME-France Performance Index employed to measure the evolution of the quality of the AROME-France forecasts. To this end, the index averages BSS values of 0.5, 2.0, 5.0 mm/6hr precipitation and 40 km/h wind gust for the 6, 12, ..." (Otherwise, one has to re-read the whole paragraph to understand what could be meant by "AROME France Performance Index" in lines 370 – 371.**

    Done

29. **Line 408: please explain why no forecasted starting during the daytime are evaluated. Have there been no differences between the 3 experiments for these forecasts?**

    Most of the SOFOG3D IOPs took place during the night and early morning. Since early forecast ranges are more sensitive to initial conditions, the simulations investigated in this study were initialized during the nighttime hours. However, meteorological situations with low flow, favorable for fog occurrence, are also favorable for persisting impact of initial conditions during the model forecast. Indeed, in old Fig. 9 (new Fig. 10), the positive impact of using initial conditions from the 3DEnVar scheme is also visible and statistically significant up to 20 h forecast ranges. Therefore, it would make sense to investigate how the forecasts initialized during daytime simulate a fog event on the following day.

30. **Line 410: better to replace "... (IOPs) were planned ... " by " ... (IOPs) were carried out ..." (?)**

    Done

31. **Lines 415 – 416: replace "... closer / higher ... better ..." by " ... the closer / the higher ... the better ..."**

    Done

32. **Line 456: "... the general better accuracy of the location of the event in the 3devi experiment."**

    Done

33. **Line 490: "location error" instead of "error location"**

    Done

34. **Lines 507 – 508: I would write: " ... forecasts (from H, the most recent, to H-15, the oldest), all valid at the same time, ..."**

    Done

35. **Line 550: "mixing the two previous elements". Maybe easier to digest: "which accounts for the forecast quality of precipitation and wind gusts"**

    Done

36. **Line 555 – 556: Why does it lead to an improvement of PEARO, given that the EDA still relies on 3DVar and hasn't changed? See comment for line 112.**

    The comment for line 112 (comment 3.9 in this document) notes that the computation of perturbed initial conditions for AROME-EPS is based on the AROME analysis. Improvements in the AROME-France deterministic analysis due to the use of 3DEnVar are thus propagated to the AROME-EPS initial conditions and forecasts. This has been made more explicit in the text.

    Due to the improved quality of the control 3DEnVar analysis, to which initial perturbations are added, significant improvements in the performance of the AROME-EPS system have been also observed

37. **Line 562: "Alpine Valley"**

    Done

38. **Line 570: "lightning"**

    Done

39. **Lines 576 – 577: Wouldn't using 5 time slots every 15 min in a 1-hour DA cycle imply that the observations in the first resp. last time slot are used twice in the DA cycle? Please comment.**

    No, the observations are not used twice, because the assimilation window is $[H - 30min, H + 30min[$: observations available at hh:30 are assimilated in the first time slot of the hh+1h analysis, not in the last time slot of the hh analysis. This is now stated in te text :

    "...4DEnVar system in order to assimilate radar, SEVIRI radiances, and surface station observations every 15 minutes in the 1-hour DA cycle (at H-30min, H-15min, H and H+15min)"

References

[1] Pierre Brousseau. Propagation de l'information observée dans le systeme d'assimilation et le modèle atmospérique AROME. Phd thesis, University of Toulouse III, France, Juillet 2012. Available at https://hal.science/tel-05494459.

[2] Marielle Amodei, Isabelle Sanchez, and Joël Stein. Verification of the french operational high-resolution model arome with the regional brier probability score. Meteorological Applications, 22(4):731–745, 2015.

Response to the reviewers of the paper "The operational 3DEnVar data assimilation scheme for the Meteo-France convective scale AROME-France"

Pierre Brousseau, Valerie Vogt, Etienne Arbogast, Maud Martet, Guillaume Thomas, Loïk Berre

February 2026

The authors would like to thank Dominik Jacques for the careful reading of the manuscript and his very constructive comments improving the manuscript quality. A point-by-point answer follows and the manuscript has been modified accordingly (blue text in the track-changes file).

1 General comments

This manuscript documents the evolution of data assimilation in the hourly cycled AROME prediction system operated and developed at Météo-France. The work presented is serious, and the results appear robust. Documenting the evolution of an operational system is valuable, as it serves as a reference both internally and for teams operating similar prediction systems at other operational centers. While the scientific quality of the work is good, the manuscript would benefit from a substantial round of revisions to improve clarity and readability. Relatively small improvements, such as the consistent use of acronyms in the text and figures, would significantly enhance readability.

1. **For example, 3DVar and EnVar could be used consistently in experiment labels instead of 3dv and 3dev, which are visually similar. Throughout the manuscript, simple reformulations would also help improve the overall flow.**

   Labels will be changed in 3DVar, EnVar and EnVarI.

2. **It is recommended that proofreading tools be used to identify and correct sentence structures and expressions that are not idiomatic in English.**

   Proofreading tools will be used when the major text modifications (see below) will be accepted.

3. **In particular, Sections 2 and 3.2 should be reworked for clarity. A major difficulty is that many systems with different configurations are described simultaneously. Although the components of the "new" operational system are all presented, relevant information is scattered among descriptions of systems used in previous validation studies. It is suggested that the emphasis be placed more clearly on the new system, while previous studies could be briefly summarized in dedicated subsections.**

   Section 2 has been deeply reworked in order to take into account these comments and to improve the clarity. The main modifications are :

   - the 3DEnVar formulation and associated notations are presented in a new sub-section 2.1,
   - the data flow is presented in a second subsection thanks to a new figure 1 and the notations introduced before,
   - a new subsection 2.4 has been added to summarize the previous studies,
   - all sensitivity experiments (and table 1) are now described at the end of the section 3.1 Methodology

   The authors took care to separate the presentation of the novel contribution (the 3DEnVar DA scheme in AROME-France) from that of Météo-France's numerical weather prediction (NWP) system, into which the 3DEnVar is integrated. Additionally, the new section 2.4 aims to clarify the technical and scientific context of the present study.

4. **Figure 1 is also particularly difficult to interpret, even with the accompanying text. Several suggestions for improving this figure are provided in the detailed comments below.**

   The authors introduce the main data flow described in this study using a new first figure (named figure 1). The old figure 1, now named figure 2, has been modified accordingly to this new figure and the reviewer comments.

For these reasons, it is recommended that the manuscript undergo a round of major revisions.

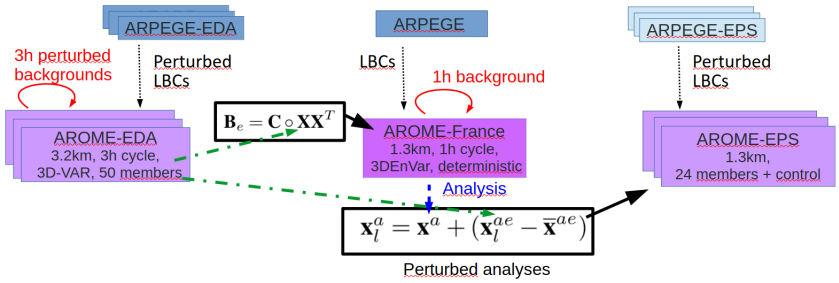


Figure 1: Schematic representation of the data exchanges between the main components of the Météo-France NWP operational system (deterministic, EDA and EPS versions) for the global model ARPEGE (blue rectangles) and the LAM AROME-France (purple rectangles). $\mathbf{B}_e$ is the ensemble background error covariance matrix, $\mathbf{C}$ the localization matrix, " $\circ$ " the Schur product and $\mathbf{X}$ the perturbation matrix provided by the AROME-EDA. $\mathbf{x}$ is a model state, $\overline{\mathbf{x}}$ the ensemble mean. Subscript l stands for an ensemble member and superscript a (resp. ae) for analysis (resp. AROME-EDA model state).

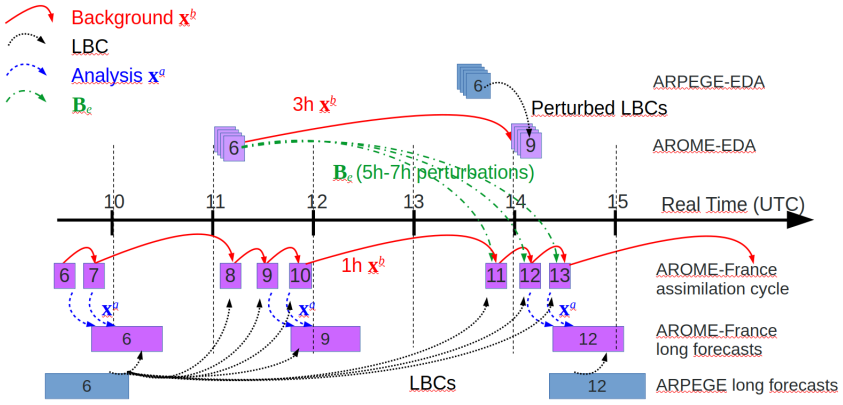


Figure 2: Temporal implementation of the operational AROME-France model from the 08 UTC to the 13 UTC analysis times (with the same pattern repeated throughout the day for other analysis times). The rectangles represent the execution of an assimilation-forecast block, each corresponding to the analysis time indicated by the number within the rectangle. These blocks are positioned along the time axis according to their actual execution time. The length of each rectangle is proportional to the wall-clock time required to complete the block on Météo-France's high-performance computing (HPC) system. Purple rectangles denote AROME runs, blue rectangles denote ARPEGE runs, and stacked rectangles represent ensemble systems. Here, $\mathbf{x}$ refers to a model state, with the superscript a (respectively b) indicating the analysis (respectively, the background state).

2 Detailed comments in the order they appear in the text

1. **Line 25 "Regarding the first limitation, numerous studies have tried to introduce some time-dependency..."**

   Would flow-dependent be more appropriate than time-dependency here? The first limitation seems more related to the nature of the error covariances than to temporal aspects. This sentence is also confusing because the second point described just above explicitly discusses the temporal dimension of the problem.

   The term "time-dependency" has been replaced by "flow-dependent features".

2. **Line 94**

   "using an original approach"

   Please be more specific about what is original in this approach.

   This point has been also raised by reviewer 1 and corrected. "Original" was employed in the text to refer to the lagged-ensemble approach used to increase the sample size of the evaluated forecasts. For each meteorological issues (winter storm, fog and HPEs), different events are investigated, and for each event, the evaluation of the 3DEnVar is done using numerous forecasts (at least 16) initialized at different analysis times in order to increase the investigation robustness.

   The wording " an original approach" is misleading and has been removed in order to avoid to add more information in the introduction.

3. **Section 2**

   Naming

   Line 102: "running a 50-member EDA"

   Line 107: "An EDA version is run"

   Presumably, these EDAs refer to ARPEGE-EPS and AROME-EPS. These acronyms are self-explanatory and should be preferred over the more obscure "PEARP" and "PEARO", which are also used in the text. They could also appear in Figure 1.

   These EDAs refer to ARPEGE-EDA and AROME-EDA and not to ARPEGE-EPS and AROME-EPS (which are ensemble prediction systems and not ensemble data assimilation systems). The use of the explicit terms ARPEGE, ARPEGE-EPS, ARPEGE-EDA, AROME-France, AROME-EPS and AROME-EDA has been generalized in the article (text and figures) in order to avoid confusion.

4. **The discussion in lines 125–145 is difficult to follow. It would help to structure this section around the new system being developed, clearly describing the start times of the different runs, and the origin of the initial conditions (ICs) and lateral boundary conditions (LBCs).**

   As mentioned above, this section has been reformulated to address this comment.

5. **Figure 1**

   This figure is difficult to understand and should be reworked. While wall-clock time is an important constraint in operational forecasting, it may not be the clearest choice for illustrating dependencies between ensemble and deterministic systems. Representing the various components in terms of forecast lead time could make the exchange of ICs, LBCs, and other information clearer.

   The authors introduce the main data flow described in this study using a new first figure (named now Fig. 1). The old Fig. 1, now named Fig. 2, has been modified accordingly to this new figure and the reviewers comment (see before). However, as the wall-clock time is an important constraint for operational NWP system and has a strong impact on the design of such systems, the authors prefer to keep it on new Fig 2. and hope that the new Fig. 1 will help to make the information clearer.

6. **Additional clarification is also needed in the discussion around line 125 and in Figure 1 itself:**

   Is "production" synonymous with forecasts?

   Yes, "production" is synonymous with "long forecasts 8 time a days" in opposition to the 1h range forecasts of the data assimilation cycle. It has been replaced by "long forecast" in the text and in new Fig. 2 (old Fig. 1).

   What do the blue and purple rectangles represent?

   Purple rectangles represent AROME run and blue ones ARPEGE runs. The information has been added in the figure's legend.

   Do stacked rectangles indicate ensemble members?

   Yes, the information has been added in the figure's legend.

7. **Line 145 refers to green arrows as "perturbed forecasts", but it is unclear which perturbations are meant. The text discusses perturbed LBCs, stochastic physics (SPP), and observation perturbations via random draws from the observation error covariance matrix for B estimation. Clarifying the different types of perturbations would be helpful.**

   The term perturbations or perturbation matrix has been explicitly defined in the new Section 2.1 (3DEnVar formulation). Additionally, it is now clearly referenced whenever it appears in the text.

   In this example, perturbed forecast is replaced by "perturbation matrix $\mathbf{X}$" :

   For example, the AROME-France analyses at 11, 12, and 13 UTC use respectively 5-, 6-, and 7-hour range perturbation matrices $\boldsymbol{X}$ (dashed-dotted green arrows), from the 06 UTC AROME-EDA analysis time, to estimate $\boldsymbol{B}_e$.

   Besides, the way model, observation, and lateral boundary condition errors are represented in AROME-EDA is now described in a dedicated paragraph of Section 2.2.

8. **Consistent use of acronyms between the figure and the text would greatly improve readability. For example, the text refers to AROME-EDA, which does not explicitly appear in the figure, while AEARP, ARO-Fr, and AEARO are not explicitly described in the text.**

   Done, see comment 1.3.

9. **Consider using the term background instead of guess.**

   Done

10. **The text refers to H–5, H–4, . . . , H, H+1, while the figure shows specific hours. Using a concrete example (e.g. 1200 UTC) could make the correspondence clearer and illustrate that the same pattern repeats at other times.**

    The text has been modified accordingly :

    "Data exchanges between the different systems are illustrated in Fig. 2. This figure specifically shows all the exchanges required to run the DA cycle of AROME-France at 11, 12, and 13 UTC, as well as the long forecasts at 12 UTC. It also depicts the sequence in which the various systems are executed and the resulting coupling choices. This execution pattern repeats throughout the day. More precisely, the 1-hour range forecasts of the DA cycle (plain red arrows) use Lateral Boundary Conditions (LBCs) from the latest ARPEGE forecast available, in an asynchronous way. For example executions at 8, 9, 10, 11, 12 and 13 UTC analysis times use LBCs from the 6 UTC ARPEGE analysis time (dotted black arrows). On the contrary, long forecasts useful for operations are launched 8 times a day (every 3 hours, for example at 9 and 12 UTC in Fig. 2) and use synchronous LBCs from ARPEGE, to benefit from the most recent LBCs, with a positive impact on the performances."

11. **Figure 1 legend : "Logical scheme of an extract of . . . " Please reformulate this sentence.**

    Done see Fig. 1 and Fig. 2 of the present document.

12. **Lines 170–180 Different pre-conditionings for different systems are discussed here. The link with the new system should be made more explicit. When referring to "the 3D-Var scheme", it is unclear whether one specific system or multiple assimilation systems are being discussed.**

    Actually, in Météo-France NWP system, only the square root form $\delta\mathbf{x} = \mathbf{B}^{1/2}\mathcal{X}$ is used, whatever the data assimilation scheme is (global 4D-Var, LAM 3D-Var or 3DEnVar). We propose to modify the text as follows :

    "which needs to be preconditioned for efficiency, through a change of variable. The square root form $\delta\boldsymbol{x} = \boldsymbol{B}^{1/2}\mathcal{X}$ which can be solved with the conjugate gradient algorithm is historically used at Météo-France."

    The reference to the 3D-Var scheme, which was used in the previous AROME-France version, and which is still used in the current operational AROME-EDA, will be precised.

13. **line 175 Minor comment: the symbol o should be used for the Schur product instead of the lowercase letter "o".**

    Done (∘)

14. **Line 181 "lead to define the settings" Please reformulate.**

    This content has been relocated to section 2.4 and reworded as follows :

    "Regarding these two articles, the following settings have been selected as a basis for the investigations described in the present study."

15. **Line 210 While precipitation can indeed be viewed as a temporal and vertical integration of model prognostic fields, using 24-hour accumulated precipitation to verify a convection-permitting, hourly cycled assimilation system seems questionable.**

As stated line 215 of the article, the 24h accumulated simulated precipitation used in section 3 are 24h accumulated precipitation simulated by the sum of the 1 h precipitation accumulation from the successive 1 h forecasts of the DA cycle over 24 h. As also explained to reviewer 1, the authors have also investigated the performances of the 1 h forecast without summation and found very consistent results. They chose to show 24 h summation of 1 h forecasts because it allows a wide range of accumulated precipitation thresholds (from 0.5 to 10 mm in 24 h) to be sampled enough to provide statistically significant results. The authors have also investigated longer forecast ranges such as 6, 12 and 18 h (without summation) of forecasts initialized at 00 and 12 UTC for all the sensitivity studies and also found very consistent results with those shown in section 3.

This information has been added in the article :

"Scores of 1 hr accumulated precipitation simulated by the 1 hr range forecasts of the DA cycle, as well as scores of longer accumulated precipitation (6 hr, 12 hr and 18 hr) have also been calculated. The results of these investigations on the sensitivity experiments are consistent with the results discussed below and will not be shown. The use of 24 hr accumulated precipitation also allows a wide range of accumulated precipitation thresholds to be sampled enough to provide statistically significant results."

16. **Line 220 (and elsewhere) The expression "strongly smaller" is awkward. Consider alternatives such as much smaller or significantly smaller.**

Done

17. **Line 225 "the differences reach only 5% in absolute value" Is Figure 2 expressed in percent? The legend refers to "relative difference", which could be interpreted in several ways.**

Yes, values in old figures 2, 3, 5, 6 and 7 (now figures 3, 4, 6, 7 and 8) are expressed in %. This has been specified in the caption of each figure.

18.a **Section 3.2 This section would be easier to follow if the experiments were described before their results are discussed.**

A short description of the sensitivity experiment characteristics has been added and the table 1 (also describing now Loc h25 and Loc h150) have been moved in section 2.1.

"Three different horizontal length scales for the horizontal localization have been first evaluated, as this parameter appears in the previous studies MO18 and MI21 to have the most visible impact. Then, different experiments have been performed to evaluate the sensitivity to the vertical localization, inflation (i.e. multiplicative factor on background error standard deviations), hybridization (i.e. use of both ensemble and climatological parts in the B matrix representation) and to the use of IAU for initializing the forecast and reducing the spin-up effects. As this section relies on the investigation of forecasts from the DA cycle, IAU is only used in the corresponding experiment to initialized the forecast with the analysis performed at H. None experiment uses IAU to update the forecast initialized at H with the H+1 analysis as explained in section 2 for long forecasts. These sensitivity experiments, described in Table 1, have been run from the 13th of September to the 15th of October 2021.

Name	Horizontal Loc.	Vertical. Loc	Inflation	Hybridization	IAU
loc h25	25 km	0.3 ln(hPa)	1	β_s=0 β_e=1	no
loc h150	150 km	0.3 ln(hPa)	1	β_s=0 β_e=1	no
Loc h25-150	25-150 km	0.3 ln(hPa)	1	β_s^2=0 β_e=1	no
Loc v0.2	25-150 km	0.2 ln(hPa)	1	β_s^2=0 β_e^2=1	no
Loc v0.4	25-150 km	0.4 ln(hPa)	1	β_s^2=0 β_e^2=1	no
Inf 0.8	25-150 km	0.3 ln(hPa)	0.8	β_s^2=0 β_e^2=1	no
Inf 1.2	25-150 km	0.3 ln(hPa)	1.2	β_s^2=0 β_e^2=1	no
Hyb	25-150 km	0.3 ln(hPa)	1	β_s^2=0.2 β_e^2=0.8	no
IAU	25-150 km	0.3 ln(hPa)	1	β_s^2=0 β_e^2=1	yes

Table 1: Summary of the experiment characteristics used in section **??**

"

18.b **It is not immediately clear whether the color red or blue indicate better performance.**

The correspondence between colors and performances is described in the caption of old figure 3 (new figure 4) but has been deliberately omitted in the caption of old figure 2 (new figure 3). In the latter,

only the correspondence between the color and the sign of the RMS difference is mentioned because it is not clear if an increase in the RMS(O-A) (i.e. when the analyses fit the observations to a lesser extent) is a degradation or an improvement, when it is simultaneously associated to a decrease of RMS(O-B) (i.e. subsequent forecasts are closer to the observations).

18.c **Using consistent labels between the figures and the text would help.**

Labels between figures and text have been made uniform.

18.d **How long were the experiments run?**

The experiments in section 3 were run during 1 month. The information is added in the text :

"These sensitivity experiments, described in Table 1, have been run from the 13th of September to the 15th of October 2021."

19. **Figure 4**

Please increase the text size for better legibility. A logarithmic scale might be appropriate here. Surface pressure tendencies are strongly influenced by the presence of convection in the domain. Plotting the envelope of minimum and maximum values in the dataset could help put the spin-up values into context.

The text size has been increased and the minimum and maximum envelope introduced for only 2 experiments (3DVar and Loc h25-150) for legibility. Attempts to use a logarithmic scale were not satisfying and we have decided to keep the linear scale.

20. **Line 260 "Spin-up can also be referred as overestimated precipitation . . . " Please provide more context, as the connection between spin-up and overestimated precipitation is not self-evident.**

Reviewer 1 had also some comments on this topic and we have decided to remove these sentences from this section. We have only added a comment on the fact that the increase of the spin-up in the 3DEnVar experiments shown on this diagnostic is not detrimental to the general behavior of the model and its performances in section 4.

21. **Line 280 "The results of these experiments are illustrated using the same diagnostics as in section 3.2. . . "**

Please clarify whether the results are shown in Figures 2 and 3, or in additional figures not shown here.

The results are shown in old figures 2 and 3 (new figures 3 and 4). This will be clearly stated in the text.

"The characteristics of the different experiments are summarized in table 1 and the results compared to the Loc h25-150 experiment are plotted in Fig. 3 and Fig. 4."

22. **Line 320 "Perturbations" Please provide more detail on the nature of these perturbations.**

As in comment 7 above, "perturbations" is replaced by *"perturbation matrix* $\boldsymbol{X}$*".*

23. **Figure 9 legend "as a fonction of the forecast range" → function**

Done (see below)

24. **Figure 10 Ensure that all text is legible; the color bar text is very small. Please also add units (mm) for the maximum values.**

Done (see below)

25. **Figure 11 Why focus on devi here? While dev and devi are similar, most earlier figures use dev.**

EnVar (previously dev) and EnVarI (previously devi) present similar behaviors and the authors have chosen to focus on EnVarI (previously devi), the final operational configuration, in order to shorten this section. EnVar (previously dev) has been removed from the new Fig. 12 (Fig. 3 in the present document) and the associated text.

26. **Section 5.3.1 It is acknowledged that sample sizes may be limited for intense precipitation events. However, the performance of a single case is largely anecdotal and does not, by itself, demonstrate the superiority of one system over another.**

The authors totally agree with this comment. That is why the section 5 uses 8 HPE cases to try to demonstrate the positive impact of the 3DEnVar scheme on HPE simulation.

In Figure 12, neither model clearly stands out as superior.

Among the 8 HPE cases presented in section 5.3 :

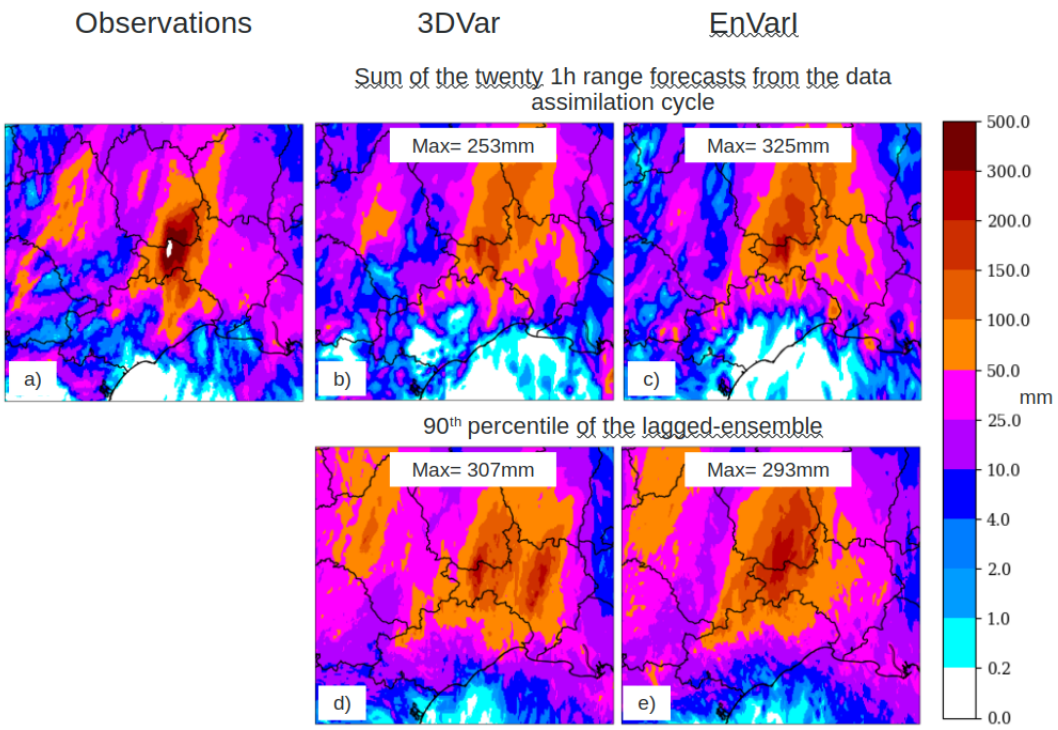


Figure 3: Total rainfall between September 19, 2020 at 04 UTC and September 20, 2020 at 00 UTC measured by the blending product (radar and rain-gauges) ANTILOPE (a) and simulated by the sum of the twenty 1-hour forecast ranges from the DA cycle (second line) and the ninetieth percentile of the lagged-ensemble built with the sixteen long-term forecasts hourly initialized from September 18, 2020 at 13 UTC to September 19, 2020 at 04 UTC (third line) for 3DVar (left), EnVar (middle) and EnVarI (right).

- the simulation of 1 case, the 19 September 2020, is clearly improved by the use of 3DEnVar and it is detailed in section 5.3.1,
- the simulation of 3 cases is neither improved nor degraded (2 October 2020, 03 October 2021 and 04 October 2021),
- the simulation of the four other cases is slightly improved thanks to the use of 3DEnVar.

About this last point, Fig. 12 is not clearly explicit but shows some improvement as stated in lines 480 to 485, 490 and 494. We agree that the improvement is not so clear as for the 19 September 2020 case on this figure, but improvements still appear on other diagnostics (not shown).

The purpose of the authors in section 5.3 was to evaluate the 3DEnVar over a maximum number of HPE cases (eight in total). We have tried to enhance the robustness of this evaluation, by using forecasts initialized at all analysis times (at least 16 forecasts for each HPE case). The main challenge was to synthesize the findings of this heavy evaluation into a concise yet comprehensive section. We are aware that drawing conclusions from a limited sample is hard, that's why we tried to draft the text in a way which ensures that the conclusions remain measured and balanced.