

Reply to anonymous referee #1

October 3, 2026

The authors would like to thank the anonymous referee #1 for their constructive comments. We have the following answers to the reviewer’s comments (“RC:” stands for reviewer comment and “AC:” stands for authors response comment).

1 Major comments

RC: *Improve the readability of the introduction, especially wrt. separation of multivariate DA and DA in integrated models you’re your discussion of different scales of previous studies around line 85 to 105 vs. the mention of the high computational burden in line 89 and 98. A better separation of the different aspects you want to point out in literature would be helpful. Also, it could lead more directly to your own study, which could mean to cut back on the discussion of multivariate DA. Moreover, I suggest to formulate your own research objectives (from line 101) more clearly. More specific comments below.*

AC: We agree that some parts of the introduction combine different aspects and that a clearer separation would help the reader. We will restructure this section of the introduction to present the ideas more clearly. In particular, we will separate the scale aspects and the computation burden aspects. We introduced the multivariate part to highlight that, while this approach may be tempting, it can also lead to mixed results, which justifies our decision not to use it. However, it is also part of the state of the art, so it may still be useful for readers to be aware of it.

The research objectives will be clarified, in line with your comments and the suggestions from the Anonymous Referee #2. In particular, we will state more explicitly that we want to evaluate the impact of the assimilation of at regional scale for future operational seasonal forecasts, and question how DA information is propagated for a groundwater model.

RC: *Performance improvements of DA: E.g. in Figure 3, Table 2, and the surrounding text: Performance improvement at assimilated sites is fine to show. However, quite expected and not that relevant in applications. I suggest to include performance at the validation wells (and adapt the text in section 3.1 accordingly). Similarly, in Figure 5 / section 3.2: For a sound evaluation of the DA improvements, I think that e.g. a cross validation through all sets of 20%/80% of observations used in assimilation/validation (or validation/assimilation) with their NIC score could be great. Results could e.g. be shown as a histogram of NIC values. In general, the text should focus more on the validation performance – this is the really exciting part of groundwater*

data assimilation in such a distributed model – can it improve it beyond the actual observation?

AC: We agree with the reviewer that the validation performance is the most important part of the study.

Running all sets of assimilated/validation wells for a cross-validation experiment would be too time-consuming and would require multiple additional simulations. However, we agree that the random selection of assimilated/validation wells could be improved.

To achieve this, we propose adding either K-fold or leave-one-out cross-validation (the latter of which has already been performed). Figure 5 would be revised to show the NIC for every well when it is used as a validation well. Figures 3 and 2 would be revised to show the results for both assimilated and validation wells. The text will be revised to reflect these modifications.

RC: *Section 4.1: As the test of the localization method is one of your main objectives: I feel it is worthwhile exploring different radii, i.e. vary the 12km. As you state yourself, this value is set based on a variographic analysis, but I assume it is uncertain whether this is the optimal value. Similar for the threshold of 0.7 in the variogram method.*

AC: We agree that the localization radius (12 km) and variogram threshold (0.7) require validation. This sensitivity analysis was part of our initial work, but not included here for simplification. The radii of 5, 8, 12 and 20 km were tested, with 12 km yielding the best NIC scores (and corresponding to an average value of the variogram range value during the variographic analysis). This essentially aligns with the findings of Li et al. (2023, cited in the manuscript), who showed that the optimal radius was derived from spatial autocorrelation analysis. In light of these results, we did not find it necessary to include the sensitivity analysis in this work. As stated on line 246, similar "manual tuning" of the threshold was performed for the variogram-based method. We started with 0.6, as Revel et al. (2019) from the original methodology, and tested 0.7 and 0.8. We finally opted for 0.7 as it was the best compromise between the NIC score and the restriction of the correlation length.

RC: *Section 3.2 / 4.3: One very relevant question in forecasting is whether the modelling system is able to beat the naïve forecasts, namely climatology and persistence. I.e. is the model (with or without DA) better at forecasting the different lead times than a simple climatology (of the observations) or assuming persistence (of the observations)?*

AC: This is a very good point. The comparative simulation against naïve forecasts was planned as part of our future work, using real seasonal forecast forcings (here, climatology refers to the climatology of forcings not observations). However, using observations for naïve forecasts as suggested could be very interesting. We will include such a comparison in Section 3.4 and discuss this in section 4.3.

2 Minor comments

RC *Why did you pick the Somme river basin as your study area? Size, data availability, good hydrogeological model, convenience, ... ? Throughout the introduction (e.g. lines 59, 71, etc) you state timeframes of positive impact of DA on groundwater level forecasting and similar. These*

timeframes must be unique to the local studies as they depend on highly variable groundwater level persistence – this should be made clear in the introduction.

AC We chose the Somme river basin because mainly because of data availability and the existence of a robust hydrogeological model that has been calibrated and used since 2002 by the French Geological Survey. While the time frames indeed are aquifer-specific, the chalk aquifer is a major aquifer in France, especially in the north of the country, so it was a good starting point for our data assimilation system. We will make this clear in the introduction as suggested.

RC *Line 75: “The authors also concluded about the importance of model’s error estimation, even if it is overestimated.” Unclear what you mean here.*

AC In the cited study, the authors concluded that the estimation of model error is a key point for the data assimilation performances and that overestimating it is still better than not considering it at all or underestimating it for accurate results. We suggest reformulating the sentence as follows: *The authors also concluded on the importance of error estimation in the model for accurate results. They stated that, if the sources of uncertainty are correctly defined, it is better to overestimate the model error than to underestimate it.*

RC *Line 84: “Multivariate DA in different component of coupled model is even more challenging (Camporese and Girotto, 2022), although possible (Thirel et al., 2010).” In my understanding, Zhang et al. (2016) also falls in this category.*

AC We agree that Zhang et al. (2016) should be added here too, we missed it here. We will add this reference to the sentence, alongside Thirel et al. (2010).

RC *Line 86: “Most of the studies cited used synthetic cases or local scale real cases. The use of DA within integrated models (coupling surface and subsurface processes) models open the way for a change of spatial scale for groundwater” What exactly do you mean with the second sentence here, or why do integrated models allow use at different (larger?) spatial scales?*

AC Actually, this is a misleading use of the term “integrated models” on our part. We suggest using “earth system models” as this better reflects the coupling of land surface models and surface/subsurface hydrological models. Such tools are often used at larger scale than “classic” groundwater models. Therefore, the use of DA in such models may facilitate a change in scale compared to local studies.

RC *Line 145: “In version 1.4, the coupling of ISBA with the hydrological models” specify that version 1.4 is the current one / the one used in this study*

AC This will be added to the revised manuscript.

RC *Line 185: “The initial hydraulic conductivity was derived from UZ thickness interpolation and pumping test data analysis (Amraoui et al., 2002).” Why do you call these “initial” – after all, you do not change these as part of your DA. Also, a word or two on setting these parameters based on this pump test data analysis vs. a potential calibration would be great. After all, especially with such a simplified model with only 1 aquifer layer, all parameter values are effective values.*

AC We will clarify this in Section 2.2.2. The term ‘initial’ refers to the parameters derived from pumping tests (Amraoui et al., 2002) and UZ thickness interpolation. While these have

not been optimized subsequently, they have been slightly modified over time through trial-error adjustments (as stated in the discussion, L509, previous to our study). We do agree that this should be stated more clearly in the section 2.2.2 and a few words could be added about the difference with an optimization as this can influence DA persistence (as discussed in section 4.3, L507-508).

RC *Line 188: The Gardénia model deserves a few words here for better understanding.*

AC Here, the Gardénia is only used as a transfer function (as stated in L189). However, we do agree that this statement is unclear to a reader.

To be more explicit, the Gardénia model is a conceptual lumped model that is typically used to calculate the water balance at basin scale. Using meteorological forcings, it employs a reservoir cascade approach to simulate infiltration through the soil, the unsaturated zone and the groundwater system. It also simulates run-off and river discharge at the basin outlet. It is fully coupled with Marthe, and usually used for surface water balance in groundwater distributed models in this context.

In our study, the infiltration and run-off are taken from the ISBA Land surface model. However, given the significant unsaturated zone of the chalk aquifer and the fact that ISBA infiltration fluxes are computed at the bottom of the root zone, we use the Gardénia model to introduce an intermediate conceptual reservoir that acts as a regulating (or delaying) factor for infiltration. This infiltration delay is computed using the exponential laws of Gardénia, and allows to convert ISBA run-off/infiltration fluxes into groundwater recharge.

Section 2.2 will be modified to reflect these explanations.

RC *Figure 1 / Appendix A1: I feel that 1 (or better even 2?) of such timeseries should be integrated into Fig. 1, as they contribute to the understanding of groundwater level dynamics in your study area.*

AC This appendix will be moved to the core material (just after figure 1), and will be described in text.

RC *Line 285ff: What are the criteria for choosing the 91 stations? How complete are the resulting timeseries?*

AC The 91 stations were selected from the 387 available in the ADES database using the following criteria: (1) temporal coverage (data from 1991 to 2024), (2) minimum frequency (12+ measurements per year), (3) database filtering for measurements tagged as “invalid”, “dry”, “pumping” and finally (4) manual quality control (visual verification of time series). On average, approximately 50 stations are assimilated per cycle, considering gaps. Most of these have daily measurements with few gaps in the simulated period (2018-2024).

RC *Line 290ff: The observation error should not only include the error of the pressure sensors in the wells. From our experience, in national public databases, there often are other errors, e.g. systematic biases introduced by wrong placement of well reference levels, etc. At least discuss what the assumption of such a small error can have on your results*

AC We acknowledge that observation errors may include additional biases such as reference level misplacement, or, as we discuss in section 4.2, the representativeness of the observation in the

model space (i.e. the difference between altimetry of the observation sites and the topography considered at the computation grid cells ; see L478-488). We will revise the text to include a mention of this placement error as a source of uncertainty in section 4.2.

RC *Line 302: Uncertainty in state estimation, i.e. model uncertainty, comes also from model structural errors and similar.*

AC We will revise line 302 to include the model structural errors as addressed in this comment.

RC *Line 305: "These fluxes were perturbed" I think it could be more clear that you mean the runoff and infiltration from ISBA here*

AC As suggested, we will make this sentence more explicit.

RC *Line 156 vs line 325: So what is the time step of the model? 1 day or 7 days?*

AC The Aquif-FR modelling platform runs at a daily time step. Therefore, ISBA timesteps are sent to hydrological models every day, and groundwater recharge is computed at the same time step. However, the groundwater state is only computed every seven days. We will make this clearer in the text.

RC *Line 334: If you select 100% of observations for DA, how do you about the validation then?*

AC Good point. It was an error in the phrasing on our part when reformulating this paragraph. The "100% assimilated" experiment simulation should not be included in the same sentence as the other experiments. We will rephrase this.

RC *Line 339: "The RMSE is then averaged for each month (from 1 to 12)" I assume you mean that the RMSE is average for each forecast horizon (1 to 6 months) across the entire year. If so, it might be more clear to write "from January to December"*

AC The average is not calculated for each forecast horizon across the entire year. Instead, the RMSE for a single forecast horizon and one initialization month is the average of the RMSE values for the same initialization month and lead time across all years of the hindcast experiment (2018-2024).

We will rephrase: *"The RMSE is averaged across all years of the hindcast experiment (2018-2024) for each forecast horizon (1 to 6 months) and each initialization month (from January to December)."*

RC *Figure 8: To be improved by adding a column that shows the OL run*

AC Figure 8 will be revised with an additional column showing OL run.